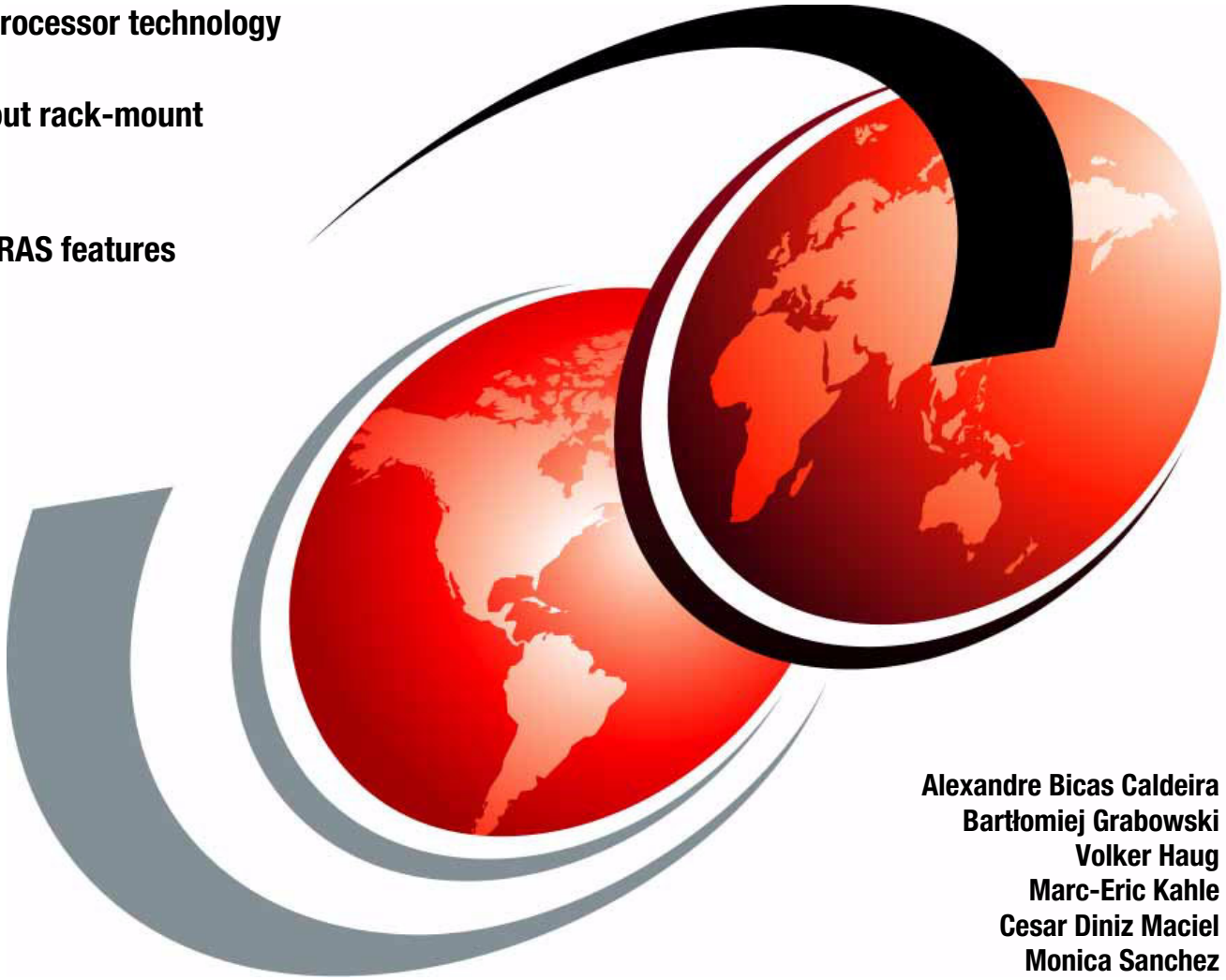
Draft Document for Review May 12, 2014 12:46 pm

REDP-5102-00

IBM Power System S822

Technical Overview and Introduction

- Outstanding performance based on POWER8 processor technology
- 2U scale-out rack-mount server
- Improved RAS features



Alexandre Bicas Caldeira
Bartłomiej Grabowski
Volker Haug
Marc-Eric Kahle
Cesar Diniz Maciel
Monica Sanchez

Redpaper

ibm.com/redbooks

International Technical Support Organization

IBM Power System S822 Technical Overview and Introduction

April 2014

Note: Before using this information and the product it supports, read the information in “Notices” on page vii.

First Edition (April 2014)

This edition applies to the IBM Power System S822 (8284-22A) server.

© Copyright International Business Machines Corporation 2014. All rights reserved.
Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Notices vii
Trademarks viii

Preface ix
Authors ix
Now you can become a published author, too! x
Comments welcome xi
Stay connected to IBM Redbooks xi

Chapter 1. General description 1
1.1 Systems overview 1
1.1.1 IBM Power S822 server 1
1.2 Operating environment 2
1.3 Physical package 3
1.4 System features 4
1.4.1 Power S822 system features 4
1.4.2 Minimum features 5
1.4.3 Power supply features 5
1.4.4 Processor module features 5
1.4.5 Memory features 6
1.4.6 PCIe slots 6
1.5 Disk and media features 7
1.6 I/O drawers for Power S822 server 9
1.6.1 EXP24S SFF Gen2-bay drawer 10
1.7 Server and virtualization management 10
1.8 System racks 12
1.8.1 IBM 7014 Model T00 rack 12
1.8.2 IBM 7014 Model T42 rack 13
1.8.3 IBM 42U Slim Rack 7965-94Y 14
1.8.4 Feature code 0551 rack 15
1.8.5 Feature code 0553 rack 15
1.8.6 Feature code ER05 rack 15
1.8.7 The AC power distribution unit and rack content 15
1.8.8 Rack-mounting rules 18
1.8.9 Useful rack additions 18
1.8.10 OEM rack 20

Chapter 2. Architecture and technical overview 23
2.1 The IBM POWER8 processor 24
2.1.1 POWER8 processor overview 25
2.1.2 POWER8 processor core 27
2.1.3 Simultaneous multithreading 28
2.1.4 Memory access 29
2.1.5 On-chip L3 cache innovation and Intelligent Cache 29
2.1.6 L4 cache and memory buffer 30
2.1.7 Hardware transactional memory 31
2.1.8 Coherent Accelerator Processor Interface 31
2.1.9 Power management and system performance 33
2.1.10 Comparison of the POWER8, POWER7+ and POWER7 processors 34

© Copyright IBM Corp. 2014. All rights reserved.

2.2 Memory subsystem . . . 34
2.2.1 Custom DIMM (CDIMM) . . . 35
2.2.2 Memory placement rules . . . 36
2.2.3 Memory bandwidth . . . 38
2.3 System Bus . . . 39
2.4 Internal I/O subsystem . . . 40
2.4.1 Slot configuration . . . 40
2.4.2 System ports . . . 43
2.5 PCI adapters . . . 43
2.5.1 PCI express . . . 43
2.5.2 LAN adapters . . . 44
2.5.3 Graphics accelerator adapters . . . 45
2.5.4 SAS adapters . . . 46
2.5.5 PCIe RAID and SSD SAS adapter . . . 46
2.5.6 Fibre Channel adapters . . . 47
2.5.7 Fibre Channel over Ethernet . . . 48
2.5.8 InfiniBand Host Channel adapter . . . 49
2.5.9 Asynchronous and USB adapters . . . 49
2.5.10 Cryptographic coprocessor . . . 50
2.5.11 FPGA Adapters . . . 50
2.6 Internal storage . . . 51
2.6.1 RAID support . . . 55
2.6.2 Easy Tier . . . 57
2.6.3 External SAS port . . . 59
2.6.4 Media bays . . . 59
2.7 External I/O subsystems . . . 60
2.8 External disk subsystems . . . 60
2.8.1 EXP24S SFF Gen2-bay drawer . . . 60
2.8.2 IBM System Storage . . . 62
2.9 Hardware Management Console (HMC) . . . 63
2.9.1 HMC code level . . . 63
2.9.2 HMC RAID 1 support . . . 64
2.9.3 HMC connectivity to the POWER8 processor-based systems . . . 65
2.9.4 High availability HMC configuration . . . 66
2.10 Operating system support . . . 67
2.10.1 AIX operating system . . . 68
2.10.2 Linux operating system . . . 69
2.10.3 Virtual I/O Server . . . 69
2.10.4 Java . . . 70
2.11 Energy management . . . 70
2.11.1 IBM EnergyScale technology . . . 70
2.11.2 On Chip Controller . . . 73
2.11.3 Energy consumption estimation . . . 74

Chapter 3. Virtualization . . . 75
3.1 POWER Hypervisor . . . 76
3.2 POWER processor modes . . . 79
3.3 Active Memory Expansion . . . 81
3.4 PowerVM . . . 85
3.4.1 PowerVM editions . . . 86
3.4.2 Logical partitions . . . 86
3.4.3 Multiple shared processor pools . . . 89
3.4.4 Virtual I/O Server . . . 90

3.4.5 PowerVM Live Partition Mobility . . . 94
3.4.6 Active Memory Sharing . . . 95
3.4.7 Active Memory Deduplication . . . 95
3.4.8 Operating system support for PowerVM . . . 96
3.4.9 Linux support . . . 97
3.4.10 PowerVM Simplification . . . 97
3.5 System Planning Tool . . . 101
3.6 IBM PowerVC . . . 103
3.7 IBM PowerVP . . . 103
3.8 VIOS 2.2.3 features . . . 105

Chapter 4. Continuous availability and manageability . . . 107
4.1 Reliability . . . 108
4.1.1 Designed for reliability . . . 108
4.1.2 Placement of components . . . 108
4.1.3 Redundant components and concurrent repair . . . 109
4.2 Availability . . . 109
4.2.1 Partition availability priority . . . 110
4.2.2 General detection and deallocation of failing components . . . 110
4.2.3 Memory protection . . . 112
4.2.4 Cache protection . . . 114
4.2.5 Special Uncorrectable Error handling . . . 114
4.2.6 PCI Enhanced Error Handling . . . 115
4.3 Serviceability . . . 116
4.3.1 Detecting . . . 116
4.3.2 Diagnosing . . . 120
4.3.3 Reporting . . . 122
4.3.4 Notifying . . . 124
4.3.5 Locating and servicing . . . 125
4.4 Manageability . . . 127
4.4.1 Service user interfaces . . . 127
4.4.2 IBM Power Systems firmware maintenance . . . 131
4.4.3 Concurrent firmware update improvements . . . 135
4.4.4 Electronic Services and Electronic Service Agent . . . 135
4.5 POWER8 RAS enhancements . . . 139
4.5.1 POWER7+ RAS features . . . 140
4.6 Power-On Reset Engine . . . 140
4.7 Operating system support for RAS features . . . 141

Related publications . . . 145
IBM Redbooks . . . 145
Other publications . . . 146
Online resources . . . 147
Help from IBM . . . 147

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

© Copyright IBM Corp. 2014. All rights reserved.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

Active Memory™
AIX®
BladeCenter®
DS8000®
Easy Tier®
Electronic Service Agent™
EnergyScale™
FlashSystem™
Focal Point™
Global Technology Services®
IBM®
IBM FlashSystem™
Micro-Partitioning®
POWER®
POWER Hypervisor™
Power Systems™
Power Systems Software™
POWER6®
POWER6+™
POWER7®
POWER7+™
POWER8™
PowerHA®
PowerLinux™
PowerPC®
PowerVM®
PowerVP™
PureFlex®
Real-time Compression™
Redbooks®
Redpaper™
Redbooks (logo) ®
RS/6000®
Storwize®
System Storage®
System x®
System z®
Tivoli®
XIV®

The following terms are trademarks of other companies:

Intel, Intel Xeon, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

LTO, Ultrium, the LTO Logo and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

Microsoft, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Other company, product, or service names may be trademarks or service marks of others.

Preface

This IBM® Redpaper™ publication is a comprehensive guide covering the IBM Power System S822 (8284-22A) server that support IBM AIX®, and Linux operating systems. The objective of this paper is to introduce the major innovative Power S822 offerings and their relevant functions:

- The new IBM POWER8™ processor, available at frequencies of 3.42 GHz, and 3.89 GHz
- Significantly strengthened cores and larger caches
- Two integrated memory controllers with improved latency and bandwidth
- Integrated I/O subsystem and hot-pluggable PCIe Gen3 I/O slots
- Improved reliability, serviceability, and availability functions
- IBM EnergyScale™ technology that provides features such as power trending, power-saving, capping of power, and thermal measurement.

This publication is for professionals who want to acquire a better understanding of IBM Power Systems™ products. The intended audience includes the following roles:

- Clients
- Sales and marketing professionals
- Technical support professionals
- IBM Business Partners
- Independent software vendors

This paper expands the current set of IBM Power Systems documentation by providing a desktop reference that offers a detailed technical description of the Power S822 system.

This paper does not replace the latest marketing materials and configuration tools. It is intended as an additional source of information that, together with existing sources, can be used to enhance your knowledge of IBM server solutions.

Authors

This paper was produced by a team of specialists from around the world working at the International Technical Support Organization, Poughkeepsie Center.

Alexandre Bicas Caldeira is a Certified IT Specialist and is a member of the Power Systems Advanced Technical Sales Support team for IBM Brazil. He holds a degree in Computer Science from the Universidade Estadual Paulista (UNESP) and a MBA in Marketing and his major areas of focus are competition, sales, and technical sales support. Alexandre has more than 14 years of experience working on IBM Systems & Technology Group Solutions and has worked also as a Business Partner on Power Systems hardware, AIX, and PowerVM® virtualization products.

Bartłomiej Grabowski is an IBM i, and PowerVM Senior Technical Specialist in DHL IT Services in the Czech Republic. He has 9 years of experience with IBM i. He holds a Bachelor's degree in Computer Science from Academy of Computer Science and Management in Bielsko-Biala. His areas of expertise include IBM i administration, PowerHA® solution based on hardware and software replication, Power Systems hardware, and

© Copyright IBM Corp. 2014. All rights reserved.

PowerVM. He is an IBM Certified Systems Expert and a coauthor of several PowerVM Redbooks®.

Volker Haug is an Open Group Certified IT Specialist within IBM Systems & Technology Group in Germany supporting Power Systems clients and Business Partners. He holds a diploma degree in Business Management from the University of Applied Studies in Stuttgart. His career includes more than 27 years of experience with Power Systems, AIX, and PowerVM virtualization; he has written several IBM Redbooks publications about Power Systems and PowerVM. Volker is a IBM POWER8 Champion and a member of the German Technical Expert Council, an affiliate of the IBM Academy of Technology.

Marc-Eric Kahle is a AIX Software specialist at the IBM Global Technology Services® in Ehningen, Germany. He also worked as a Power Systems Hardware Support specialist in the RS/6000®, Power Systems, and AIX fields since 1993. He has worked at IBM Germany since 1987. His areas of expertise include Power Systems hardware and he is an AIX certified specialist. He has participated in the development of seven other IBM Redbooks publications.

Cesar Diniz Maciel is an Executive IT Specialist with IBM in the United States. He joined IBM in 1996 as Presales Technical Support for the IBM RS/6000 family of UNIX servers in Brazil, and came to IBM United States in 2005. He is part of the Global Techline team, working on presales consulting for Latin America. He holds a degree in Electrical Engineering from Universidade Federal de Minas Gerais (UFMG) in Brazil. His areas of expertise include Power Systems, AIX, and IBM POWER® Virtualization. He has written extensively on Power Systems and related products. This is his eighth ITSO residency.

Monica Sanchez is an Advisory Software Engineer with more than 13 years of experience in AIX and Power Systems support. Her areas of expertise include AIX, HMC, and networking. She holds a degree in Computer Science from Texas A&M University and is currently part of the Power HMC Product Engineering team, providing level 2 support for the IBM Power Systems Hardware Management Console.

The project that produced this publication was managed by:

Scott Vetter
Executive Project Manager, PMP

Thanks to the following people for their contributions to this project:

Tamikia Barrow, Terry Brennan, Ron Brown, Carl Burnett, Charlie Burns, Bill Buros, Johnathan Dement, Dan Dumarot, Jessica Erber-Stark, Medha D. Fox, Ben Gibbs, Chuck Graham, Kris Kendall, Mark Kressin, Karen Lawrence, Woodrow Lemcke, Edward Liu, Pat Mackall, Michael MacPherson, Ricardo Marin Matinata, Bruce Mealey, Dirk_Michel, Michael J. Mueller, Thoi Ngyuen, Mark Olson, Bret Olszewski, Kanisha Patel, Rajendra Patel, Vani Ramagiri, Pat O'Rourke, Armin Roell, Todd Rosedahl, Jeff Scheel, Guillermo J. Silva, Jeff Stuecheli, Claire Toth, Madeline Vega, Julian Wang
IBM

Louis Bellanger
Bull

Yuki Taniguchi
Hitachi

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our papers to be as helpful as possible. Send us your comments about this paper or other IBM Redbooks publications in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:

  ibm.com/redbooks
- Send your comments in an email to:

  redbooks@us.ibm.com
- Mail your comments to:

  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks
- Follow us on Twitter:

  http://twitter.com/ibmredbooks
- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806
- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm
- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

1

General description

The IBM Power System S822 (8284-22A) server uses the latest POWER8 processor technology that is designed to deliver unprecedented performance, scalability, reliability, and manageability for demanding commercial workloads.

This server brings together business transaction processing with infrastructure for social and mobile solutions in UNIX and Linux operating environments. Based on POWER8 processors, this server delivers two times the throughput of Intel based offerings for comparable workloads and provides superior economics for scale-out deployments.

The IBM Power S822 server for existing customers is designed to put data to work. With a strong commitment to optimizing AIX workloads, this system delivers better performance than the prior generation of systems and offers unmatched price and performance value for integrated Linux applications.

The one or two socket server provide the ideal foundation for private and public cloud infrastructure. The Power S822 is available in 6-core, 10-core, 12-core, and 20-core configurations and supports new I/O capabilities, including Coherent Accelerator Processor Interface (CAPI) accelerators, higher internal disk and SSD capacities, and hot plug PCIe Gen3 slots. The highly secure architecture provides a stable database and middleware platform for efficient deployment of business processing applications.

1.1 Systems overview

The following sections provide detailed information about the Power S822 system.

1.1.1 IBM Power S822 server

The Power S822 (8284-22A) server offers greater performance per core and per socket with POWER8 processors, new I/O capabilities, higher internal storage and PCIe capacities and performance, the capability to support CAPI accelerator devices, and greater reliability, availability, and serviceability (RAS), including hot-plug PCIe capability.

© Copyright IBM Corp. 2014. All rights reserved.

The Power S822 server supports a maximum of 16 DDR3 CDIMM slots. Memory features supported are 16 GB, 32 GB, and 64 GB and run at 1600 MHz, allowing for a maximum system memory of 1024 GB.

The high data transfer rates that are offered by the PCIe Gen3 slots can allow higher I/O performance or consolidation of the I/O demands on to fewer adapters running at higher rates. The result is better system performance at a lower cost when I/O demands are high.

The Power S822 offers three storage backplane options, providing a great deal of flexibility and capability. The high performance SAS controller provides RAID-0, RAID-5, RAID-6 and RAID-10 support for either HDD or SSD. One of the following three options are available for choice:

- Storage backplane with SFF-3 bays and a DVD bay
- Split backplane functionality: This feature modifies the base storage backplane cabling and adds a second, high performance SAS controller. The existing twelve SFF-3 SAS bays are cabled to be split into two sets of six bays, each with one SAS controller. Both SAS controllers are located integrated slots and do not use a PCIe slot.
- Storage Backplane with eight SFF-3 bays, 1.8-inch SSD cage, a DVD bay, and a dual IOA with Write Cache and IBM Easy Tier® functionality:

  The IBM Easy Tier function is provided with the dual I/O adapter (IOA), This function is implemented just within the integrated Power Systems SAS controllers, the integrated SAS bays, and, optionally, in an EXP24S I/O drawer. Hot data is automatically moved to SSD, and cold data is automatically moved to HDD in an AIX, Linux, or VIOS environment.

The Active Memory™ Expansion feature enables memory expansion on the system. Using compression/decompression of memory, content can effectively expand the maximum memory capacity, providing additional server workload capacity and performance.

IBM EnergyScale technology provides features, such as power trending, power-saving, capping of power, and thermal measurement.

Figure 1-1 shows the Power S822 server.

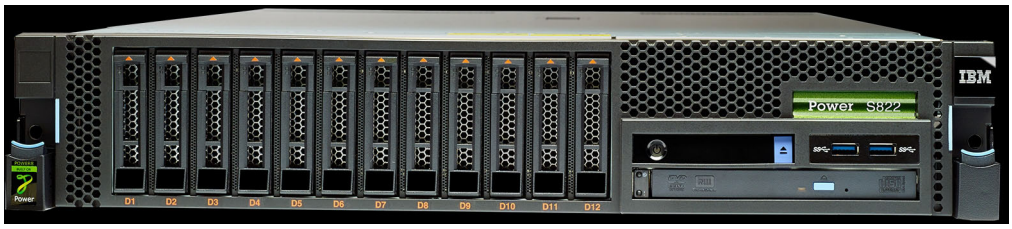


Figure 1-1 Front view of the Power S822

1.2 Operating environment

Table 1-1 on page 3 lists the operating environment specifications for the Power S822 server.

Table 1-1 Operating environment for Power S822

Power S822 operating environment		
Description	**Operating**	**Non-operating**
Temperature	Allowable: 5 - 35[a] degrees C (41 - 95 degrees F) Recommended: 18 to 27 degrees C (64 to 80 degrees F)	5 - 45 degrees C (41 - 113 degrees F)
Relative humidity	8 - 80%	8 - 80%
Maximum dew point	28 degrees C (84 degrees F)	N/A
Operating voltage	200 - 240 V AC	N/A
Operating frequency	47 or 63 Hz	N/A
Power consumption	1810 Watts maximum	N/A
Power source loading	1.88 kVA maximum	N/A
Thermal output	6176 BTU/hour maximum	N/A
Maximum altitude	3,050 m (10,000 ft)	N/A
Noise level and sound power	6.7 bels operating; 6.7 bels idling	N/A

a. Heavy workloads may see some performance degradation above 35C if internal temperatures trigger a CPU clock reduction.

Tip: The maximum measured value is expected from a fully populated server under an intensive workload. The maximum measured value also accounts for component tolerance and operating conditions that are not ideal. Power consumption and heat load vary greatly by server configuration and utilization. Use the IBM Systems Energy Estimator to obtain a heat output estimate based on a specific configuration, available at the following website:

http://www-912.ibm.com/see/EnergyEstimator

1.3 Physical package

Table 1-2 shows the physical dimensions of the Power S822 chassis. The server is available only in a rack-mounted form factor and takes 2U (2 EIA units) of rack space.

Table 1-2 Physical dimensions

Dimension	Power S822 (8284-22A)
Width	443 mm (17.5 in)
Depth	755 mm (29.7 in)
Height	87 mm (3.5 in)
Weight (maximum configuration)	28.6 kg (63 lbs)

Figure 1-2 shows the rear view of a Power S822 system.

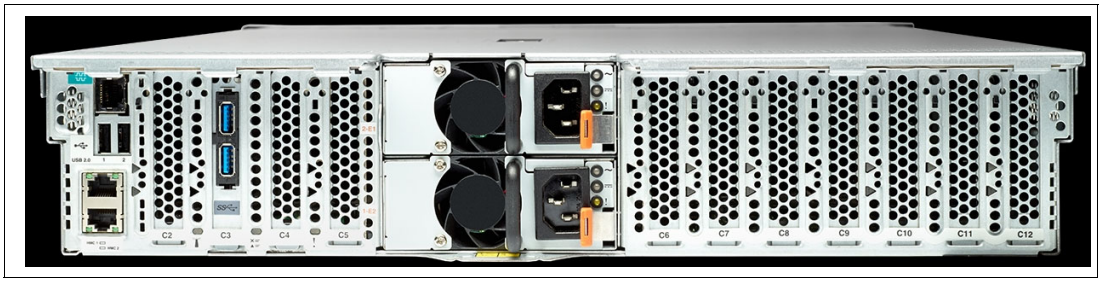


Figure 1-2 Rear view of a Power S822 system

1.4 System features

The system chassis contains one or two processor modules. Each POWER8 processor module is either 6-core, or 10-core and has a 64-bit architecture, up to 512 KB of L2 cache per core, and up to 8 MB of L3 cache per core.

1.4.1 Power S822 system features

This summary describes the standard features of the Power S822:

- Rack-mount (2U) chassis
- Single or dual processor module:
  - 6-core 3.89 GHz processor module
  - 10-core 3.42 GHz processor module
- Up to 1024 GB of 1600 MHz DDR3 ECC memory
- Choice of two storage features:
  - Choice one:
    - Twelve SFF bays, one DVD bay, one integrated SAS controller without cache, and JBOD RAID 0, 5, 6, or 10
    - Optionally, split the twelve SFF-3 bays and add a second integrated SAS controller without cache.
  - Choice two:
    - Eight SFF-3 bays, one DVD bay, a pair of integrated SAS controllers with cache, RAID 0, 5, 6, 10, 5T2, 6T2, and 10T2.
    - A 6-bay, 1.8-inch SSD cage with dual IOA.
    - Optionally, attach an EXP24S SAS HDD/SSD expansion drawer to the dual IOA.
- Hot swap PCIe Gen3 LP slots
  - Nine slots with two processor DCMs: four x16 slots and five x8 slots
  - Six slots with one processor DCM: two x16 slots and four x8 slots

Note:

- One of the x8 PCIe slots on the Power S822 server is used for a PCIe2 LP 4-port 1 Gb Ethernet Adapter (#5260)
- One fewer PCIe slot is available with the dual IOA storage backplane feature EJ0U

- One DVD-RAM drive
- Integrated features:
  - Service processor
  - EnergyScale technology
  - Hot-swap and redundant cooling
  - USB 3.0 and 2.0 ports
  - Two HMC ports
  - One system port with RJ45 connector
- Redundant, 1400W hot-swap power supplies

1.4.2 Minimum features

The minimum Power S822 initial order must include a processor module, processor activations, 32 GB of memory, memory activations, a storage backplane, one HDD or SDD, a PCIe2 LP 4-port 1Gb Ethernet adapter (#5260), two power supplies and line cords, an operating system indicator, a cover set indicator, and a Language Group Specify. AIX or Linux is the primary operating system.

Note: IBM i is not supported on the Power System S822.

1.4.3 Power supply features

Two redundant 1400 Watt 200-240 Volt power supplies (#EB2M) are supported on the Power S822 server.

The server will continue to function with one working power supply. A failed power supply can be hot-swapped but must remain in the system until the replacement power supply is available for exchange.

1.4.4 Processor module features

A maximum of two processors cards of either six processor cores (#EPX1) or ten processor cores (#EPXD) are allowed. All processor cores must be activated unless they are factory deconfigured by using feature code 2319.

The following listing defines the allowed quantities of processor activation entitlements:

- One 6-core, 3.89 GHz processor (#EPX1) requires that six processor activation codes be ordered. A maximum of six processor activation code features (#EPY1) is required.
- Two 6-core, 3.89 GHz processors (#EPX1) require that twelve processor activation codes be ordered. A maximum of twelve processor activation code features (#EPY1) is required.
- One 10-core, 3.42 GHz processor (#EPXD) requires that ten processor activation codes be ordered. A maximum of ten processor activation code features (#EPY1) is required.

- Two 10-core, 3.42 GHz processors (#EPXD) requires that twenty processor activation codes be ordered. A maximum of twenty processor activation code features (#EPY1) is required.

Table 1-3 summarizes the processor features that are available for the Power S822.

Table 1-3 Processor features for the Power S822

Feature code	Processor module description
EPX1	6-core 3.89 GHz POWER8 processor card
EPXD	10-core 3.42 GHz POWER8 processor card

1.4.5 Memory features

A minimum 32 GB of memory is required on the Power S822 system. Memory upgrades require memory installed in pairs. The minimum memory is two 16 GB, 1600 MHz DDR3 memory modules (#EM83).

Plans for future memory upgrades should be taken into account when deciding which memory feature size to use at the time of initial system order.

Table 1-4 lists memory features that are available on the Power S822 system.

Table 1-4 Summary of memory features

Feature code	DIMM capacity	Access rate	Maximum quantity
EM83	16 GB	1600 MHz	16
EM84	32 GB	1600 MHz	16
EM85	64 GB	1600 MHz	16

1.4.6 PCIe slots

The Power S822 has up to nine PCIe hot plug Gen3 slots, providing excellent configuration flexibility and expandability. For future usage of even more PCIe slots, there is a Statement of Direction for an I/O drawer with PCIe slots available.

With two POWER8 processor dual-chip modules (DCM), a maximum of nine PCIe Gen3 slots are available. Four are x16 LP slots, and five are x8 Gen3 LP slots.

With one POWER8 processor DCM, a maximum of six PCIe Gen3 slots are available. Two are x16 LP slots, and four are x8 Gen3 LP slots.

The x16 slots can provide up to twice the bandwidth of x8 slots because they offer twice as many PCIe lanes. PCIe Gen3 slots can support up to twice the bandwidth of a PCIe Gen2 slot and up to four times the band-width of a PCIe Gen1 slot, assuming an equivalent number of PCIe lanes.

Note: One of the x8 PCIe slots is used for a PCIe2 LP 4-port 1Gb Ethernet Adapter (#5260).

The new servers are smarter about energy efficiency for cooling PCIe adapter environment. They sense which IBM PCIe adapters are installed in their PCIe slots and if an adapter

requires higher levels of cooling, they automatically speed up fans to increase airflow across the PCIe adapters.

1.5 Disk and media features

Three backplane options are available for the Power S822 and provide a great deal of flexibility and capability. One of the three options must be configured:

1. Storage Backplane with 12 SFF-3 bays and one DVD bay (#EJ0T)
2. Storage Backplane with 12 SFF-3 bays and one DVD bay (#EJ0T) and feature EJ0V provides split backplane functionality
3. Storage Backplane with 8 SFF-3 bays, 1.8-inch SSD cage bays, one DVD bay, dual integrated SAS controllers with write cache and Easy Tier functionality (#EJ0U)

Each of the three backplane options provides SFF-3 SAS bays in the system unit. These 2.5-inch or small form factor (SFF) SAS bays can contain SAS drives (HDD or SSD) mounted on a Gen3 tray or carrier. Thus the drives are designated SFF-3. SFF-1 or SFF-2 drives do not fit in an SFF-3 bay. All SFF-3 bays support concurrent maintenance or *hot plug* capability.

In addition to supporting HDDs and SSDs in the SFF-3 SAS bays of the Power S822, the storage backplane feature EJ0U supports a mandatory 6-bay, 1.8-inch SSD Cage (#EJTL). All six bays are accessed by both of the integrated SAS controllers. The bays support concurrent maintenance (hot plug). The SSD 1.8-inch drive such as the 387 GB capacity feature ES16 (AIX, Linux) is supported.

The high performance SAS controllers provide RAID-0, RAID-5, RAID-6 and RAID-10 support. The dual SAS controllers can automatically move hot data to attached SSD and cold data to attached HDD for AIX, and Linux environments using the Easy Tier functionality.

Table 1-5 shows the available disk drive feature codes that can be installed in the Power S822.

Table 1-5 Disk drive feature code description for Power S822.

Feature code	CCIN	Description	Max	OS support
1917	19A3	146 GB 15 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	576	AIX, Linux
ESDT		146 GB 15 K RPM SAS SFF-3 Disk Drive (AIX/Linux)	12	AIX, Linux
1925	31D0	300 GB 10 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	576	AIX, Linux
ESDR		300 GB 10 K RPM SAS SFF-3 Disk Drive (AIX/Linux)	12	AIX, Linux
ESEZ		300 GB 15 K RPM SAS SFF-2 4 K Block - 4096 Disk Drive	576	AIX, Linux
1953	19B1	300 GB 15 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	576	AIX, Linux
ESFB		300 GB 15 K RPM SAS SFF-3 4 K Block - 4096 Disk Drive	12	AIX, Linux

Feature code	CCIN	Description	Max	OS support
ESDB		300 GB 15 K RPM SAS SFF-3 Disk Drive (AIX/Linux)	12	AIX, Linux
ES16		387 GB 1.8" SSD for AIX/Linux	6	AIX, Linux
ES19		387 GB SFF-2 SSD for AIX/Linux	288	AIX, Linux
ES0C	54BE	387 GB SFF-2 SSD for AIX/Linux with eMLC	288	AIX, Linux
ES0L		387 GB SFF-3 SSD for AIX/Linux	12	AIX, Linux
1964	19B3	600 GB 10 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	576	AIX, Linux
ESD5		600 GB 10 K RPM SAS SFF-3 Disk Drive (AIX/Linux)	12	AIX, Linux
ESFP		600 GB 15 K RPM SAS SFF-2 4 K Block - 4096 Disk Drive	576	AIX, Linux
ESFF		600 GB 15 K RPM SAS SFF-3 4 K Block - 4096 Disk Drive	12	AIX, Linux
ES0G	54BE	775 GB SFF-2 SSD for AIX/Linux	288	AIX, Linux
ES0N		775 GB SFF-3 SSD for AIX/Linux	12	AIX, Linux
1752	19B4	900 GB 10 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	576	AIX, Linux
ESD3		1.2 TB 10 K RPM SAS SFF-2 Disk Drive (AIX/Linux)	576	AIX, Linux
ESD9		1.2 TB 10 K RPM SAS SFF-3 Disk Drive (AIX/Linux)	12	AIX, Linux
EQ52	54BE	Quantity 150 of #1752 (900 GB SAS SFF-2 disk)	3	AIX, Linux
1866	2BC1	Quantity 150 of #1917 (146 GB SAS SFF-2 disk)	3	AIX, Linux
1869	2BC1	Quantity 150 of #1925 (300 GB SAS SFF-2 disk)	3	AIX, Linux
1929	31D0	Quantity 150 of #1953 (300 GB SAS SFF-2 disk)	3	AIX, Linux
1818	2BC6	Quantity 150 of #1964 (600 GB SAS SFF-2 disk)	3	AIX, Linux
EQ0G	54BE	Quantity 150 of #ES0G (775 GB SSD SFF-2)	2	AIX, Linux
EQ19	54BE	Quantity 150 of #ES19 (387 GB SSD SFF-2)	1	AIX, Linux
EQD3	54BE	Quantity 150 of #ESD3 (1.2 TB 10 K SFF-2)	3	AIX, Linux
EQEZ	54BE	Quantity 150 of #ESEZ (300 GB SFF-2)	3	AIX, Linux
EQFP	54BE	Quantity 150 of #ESFP (600 GB SFF-2)	3	AIX, Linux
EQ0C	54BE	Quantity of 150 #ES0C	1	AIX, Linux

Included in the feature EJ0T or EJ0U backplanes is a slimline media bay that can optionally house a SATA DVD-RAM (#5771). The DVD drive is run by the integrated SAS controllers, and a separate PCIe adapter is not required.

The Power S822 supports the RDX USB External Docking Station for Removable Disk Cartridge (#EU04). The USB External Docking Station accommodates RDX removable disk cartridge of any capacity. The disks are in a protective rugged cartridge enclosure that plug into the docking station. The docking station holds one removable rugged disk drive/cartridge at a time. The rugged removable disk cartridge and docking station backs up similar to tape drive. This can be an excellent alternative to DAT72, DAT160, 8mm, and VXA-2 and VXA-320 tapes.

Table 1-6 shows the available media device feature codes for Power S822.

Table 1-6 Media device feature code description for Power S822

Feature code	Description
5771	SATA Slimline DVD-RAM Drive
EU04	RDX USB External Docking Station for Removable Disk Cartridge

SCSI disks are not supported in the Power S822 disk bays. Also, because there is no PCIe LP SCSI adapter available, you cannot attach existing SCSI disk subsystems.

If you need more disks than are available with the internal disk bays, you can attach additional external disk subsystems. For more information about the available external disk subsystems, see 2.8, “External disk subsystems” on page 60.

For more information about the internal disk features, see 2.6, “Internal storage” on page 51.

1.6 I/O drawers for Power S822 server

At the time of writing there is no PCIe I/O drawer supported on the Power S822.

Statement of direction: IBM plans to introduce an PCIe I/O drawer which will expand the number of PCIe Gen3/Gen2 slots available on POWER8 processor-based Power Systems.

Customers migrating from earlier generation Power Systems servers may have been using I/O drawers such as the GX++ attached feature 5802 or 5877 PCIe 12X I/O Drawers with PCIe Gen1 slots. Though most of the disk drives in the feature 5802 can be converted and moved to the feature 5887 EXP24S drawer, the feature 5802 and 5877 drawers are not supported on this newer Power System built with POWER8 based technology.

Similarly, the GX++ attached EXP30 Ultra SSD Drawer (#EDR1 or #5888) is not supported. Also the 3.5-inch-based feature 5886 EXP12S SAS Disk Drawer and feature 5786 EXP24 SCSI Disk Drawer are not supported.

IBM offers the IBM System Storage® 7226 Model 1U3 Multi-Media Enclosure that can hold one or more DVDs, tape drive, or RDX docking stations. For more information about the multimedia drawer, see “IBM System Storage 7226 Model 1U3 Multi-Media Enclosure” on page 18.

1.6.1 EXP24S SFF Gen2-bay drawer

If you need more disks than are available with the internal disk bays, you can attach additional external disk subsystems such as the EXP24S SAS HDD/SSD expansion drawer (#5887). The EXP24S SFF Gen2-bay drawer is an expansion drawer supporting up to 24 2.5-inch hot-swap SFF SAS HDDs on IBM POWER6®, IBM POWER6+™, POWER7®, POWER7+™, or POWER8 servers in 2U of 19-inch rack space. The EXP24S bays are controlled by SAS adapters or controllers attached to the I/O drawer by SAS X or Y cables.

The EXP24S drawer is attached to SAS ports on either a PCIe SAS adapter located in the server or to the SAS ports on the rear of the server. Two SAS ports on the rear of the server are enabled with the expanded-function storage backplane with dual IOA support (#EJ0U).

A maximum of 14 EXP24S drawers are supported on the Power S822.

The SFF bays of the EXP24S differ from the SFF bays of the POWER8 system units. The EXP24S uses Gen2 or SFF-2 SAS drives that physically do not fit in the SFF-3 bays of the POWER8 system unit.

The EXP24S includes redundant AC power supplies and two power cords.

Figure 1-3 shows EXP24S SFF drawer.

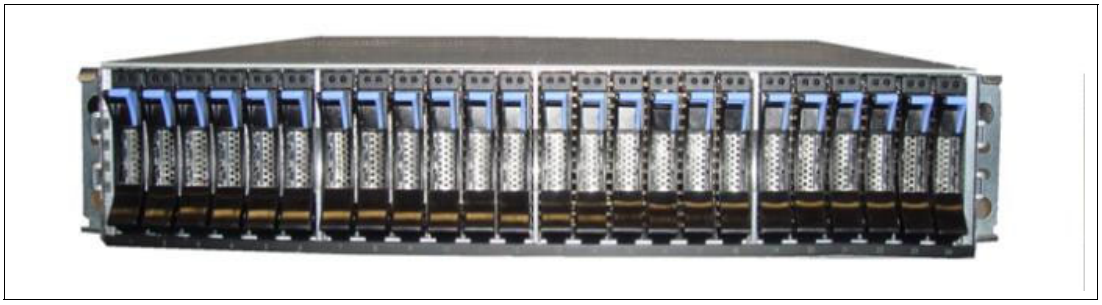


Figure 1-3 EXP24S SFF drawer

1.7 Server and virtualization management

If you want to implement logical partitions, a Hardware Management Console (HMC) or the Integrated Virtualization Manager (IVM) is required to manage the Power S822 server. In general, multiple IBM POWER6, POWER6+, POWER7, POWER7+, and POWER8 processor-based servers can be supported by a single HMC.

Remember: If you do not use an HMC or IVM, the Power S822 runs in full system partition mode. That means that a single partition owns all the server resources, and only one operating system can be installed.

If an HMC is used to manage the Power S822, the HMC must be a rack-mount CR5 or later, or deskside C08 or later.

In April 2014, IBM announced a new HMC model, machine type 7042-CR8. Hardware features on the CR8 model include a second disk drive (#1998) for RAID 1 data mirroring, and the option of a redundant power supply. If you prefer not to have RAID 1 enabled on the HMC, you can override it in the ordering system and remove the additional hard disk drive

(HDD) from the order. RAID 1 is also offered on the 7042-CR6, 7042-CR7, and 7042-CR8 as an MES upgrade option.

At the time of writing, the latest version of HMC code is V8R8.1.0. This code level also includes the LPAR function support, which allows the HMC to manage more LPARs per processor core. A core can be partitioned in up to 20 LPARs (0.05 of a core).

Several HMC models are supported to manage POWER8 processor-based systems. The models 7042-CR7 and 7042-CR8 are the only HMC available for ordering at the time of writing, but you can also use one of the withdrawn models listed in Table 1-7.

Table 1-7 HMC models that support POWER8 processor technology-based servers

Type-model	Availability	Description
7042-C08	Withdrawn	IBM 7042 Model C08 Deskside Hardware Management Console
7042-CR5	Withdrawn	IBM 7042 Model CR5 Rack-Mounted Hardware Management Console
7042-CR6	Withdrawn	IBM 7042 Model CR6 Rack mounted Hardware Management Console
7042-CR7	Available	IBM 7042 Model CR7 Rack mounted Hardware Management Console
7042-CR8	Available	IBM 7042 Model CR8 Rack mounted Hardware Management Console

The IBM POWER8 processor-based Power S822 server require HMC V8R8.1.0. The HMC V8.8.1.0 code level supports the following improvements and abilities:

- Support for managing the IBM Power System S822, S814, and S824, and IBM Power S822L systems
- Support for the new HMC model 7042-CR8
- Support for PowerVM functions such as the new HMC GUI interface for VIOS management
- GUI for HMCs Performance and Capacity Monitoring function
- A new HMC command to initiate a remote restart operation. This removes the requirement of VMControl for the PowerVM Remote Restart function.
- Management of IBM BladeCenter® Power Blade servers. This includes support for dual VIOS, Live Partition Mobility between blades and rack servers, and management of both blades and rack servers from a single management console.
- Updated browser support: The HMC now supports Mozilla Firefox ESR 17 and 24, Microsoft Internet Explorer 9 through 11, and Chrome 31.

Tip: You can download or order the latest HMC code from the Fix Central website:

`http://www.ibm.com/support/fixcentral`

If you are attaching an HMC to a new server or adding function to an existing server that requires a firmware update, the HMC machine code may need to be updated because HMC code must always be equal to or higher than the managed server's firmware. Access to firmware and machine code updates is conditioned on entitlement and license validation in accordance with IBM policy and practice. IBM may verify entitlement through customer number, serial number, electronic restrictions, or any other means or methods employed by IBM at its discretion.

1.8 System racks

The Power S822 is designed to mount in the 36U 7014-T00 (#0551), the 42U 7014-T42 (#0553), or the IBM 42U Slim Rack (7965-94Y) rack. These racks are built to the 19-inch EIA 310D standard.

Order information: The racking approach for the initial order must be either a 7014-T00, 7014-T42, or 7965-94Y. If an additional rack is required for I/O expansion drawers as an MES to an existing system, either a feature 0551, 0553, or ER05 rack must be ordered.

If a system is to be installed in a rack or cabinet that is not IBM, ensure that the rack meets the requirements described in 1.8.10, “OEM rack” on page 21.

Responsibility: The client is responsible for ensuring that the installation of the drawer in the preferred rack or cabinet results in a configuration that is stable, serviceable, safe, and compatible with the drawer requirements for power, cooling, cable management, weight, and rail security.

1.8.1 IBM 7014 Model T00 rack

The 1.8-meter (71-inch) model T00 is compatible with past and present IBM Power Systems servers. The features of the T00 rack are as follows:

- Has 36U (EIA units) of usable space.
- Has optional removable side panels.
- Has optional side-to-side mounting hardware for joining multiple racks.
- Has increased power distribution and weight capacity.
- Supports both AC and DC configurations.
- Up to four power distribution units (PDUs) can be mounted in the PDU bays (see Figure 1-5 on page 16), but others can fit inside the rack. See 1.8.7, “The AC power distribution unit and rack content” on page 15.
- For the T00 rack three door options are available:
  - Front Door for 1.8 m Rack (#6068)

    This feature provides an attractive black full height rack door. The door is steel, with a perforated flat front surface. The perforation pattern extends from the bottom to the top of the door to enhance ventilation and provide some visibility into the rack.
  - A 1.8 m Rack Acoustic Door (#6248)

    This feature provides a front and rear rack door designed to reduce acoustic sound levels in a general business environment.
  - A 1.8 m Rack Trim Kit (#6263)

    If no front door will be used in the rack, this feature provides a decorative trim kit for the front.
- Ruggedized Rack Feature

  For enhanced rigidity and stability of the rack, the optional Ruggedized Rack Feature (#6080) provides additional hardware that reinforces the rack and anchors it to the floor. This hardware is designed primarily for use in locations where earthquakes are a concern. The feature includes a large steel brace or truss that bolts into the rear of the rack.

It is hinged on the left side so it can swing out of the way for easy access to the rack drawers when necessary. The Ruggedized Rack Feature also includes hardware for bolting the rack to a concrete floor or a similar surface, and bolt-in steel filler panels for any unoccupied spaces in the rack.

- Weights are as follows:
  - T00 base empty rack: 244 kg (535 lb.)
  - T00 full rack: 816 kg (1795 lb.)
  - Maximum weight of drawers is 572 kg (1260 lb.)
  - Maximum weight of drawers in a zone 4 earthquake environment is 490 kg (1080 lb.). This number equates to 13.6 kg (30 lb.) per EIA.

Important: If additional weight is added to the top of the rack, for example adding #6117, the 490 kg (1080 lb.) must be reduced by the weight of the addition. As an example, #6117 weighs approximately 45 kg (100 lb.) so the new maximum weight of drawers that the rack can support in a zone 4 earthquake environment is 445 kg (980 lb.). In the zone 4 earthquake environment, the rack must be configured starting with the heavier drawers at the bottom of the rack.

1.8.2 IBM 7014 Model T42 rack

The 2.0-meter (79.3-inch) Model T42 addresses the client requirement for a tall enclosure to house the maximum amount of equipment in the smallest possible floor space. The following features are for the model T42 rack (which differ from the model T00):

- The T42 rack has 42U (EIA units) of usable space (6U of additional space).
- The model T42 supports AC power only.
- Weights are as follows:
  - T42 base empty rack: 261 kg (575 lb.)
  - T42 full rack: 930 kg (2045 lb.)

The available door options for T42 rack are shown in Figure 1-4.

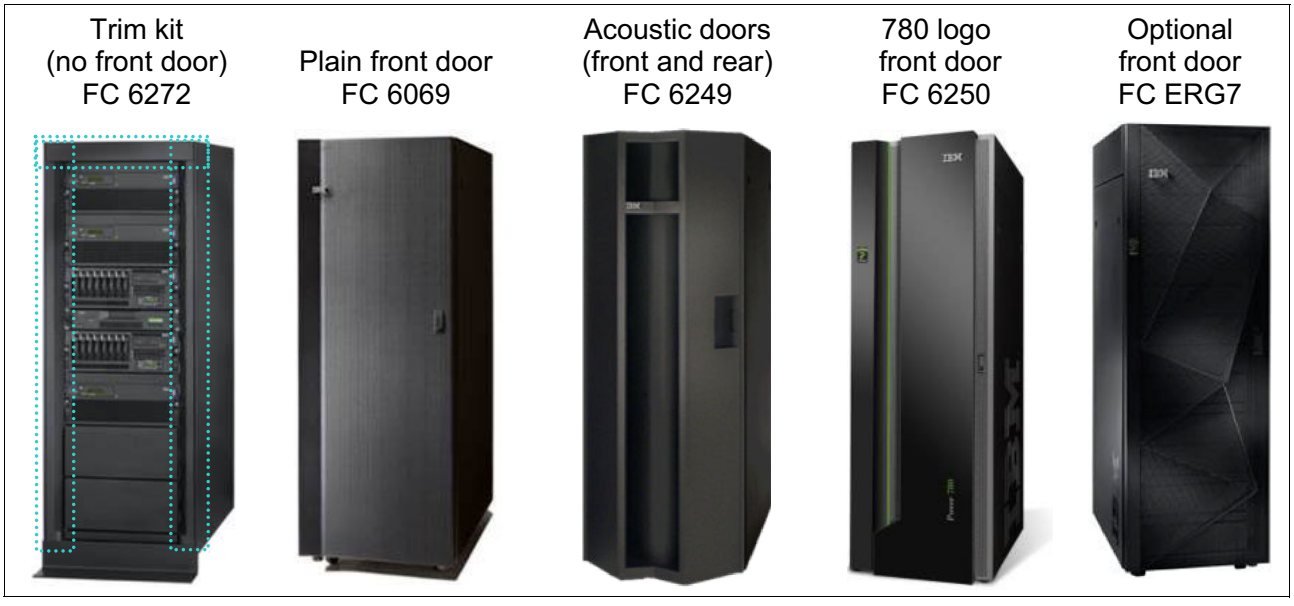


Figure 1-4 Door options for the T42 rack

- The 2.0 m Rack Trim Kit (#6272) is used, if no front door is used in the rack.

- The Front Door for a 2.0 m Rack (#6069) is made of steel, with a perforated flat front surface. The perforation pattern extends from the bottom to the top of the door to enhance ventilation and provide some visibility into the rack. This door is non acoustic and has a depth of about 25 mm (1 in).
- The 2.0 m Rack Acoustic Door (#6249) consists of a front and rear door to reduce noise by approximately 6 dB(A). It has a depth of approximately 191 mm (7.5 in).
- The High-End Appearance Front Door (#6250) provides a front rack door with a field-installed Power 780 logo indicating that the rack contains a Power 780 system. The door is not acoustic and has a depth of about 90 mm (3.5 in).

High end: For the High-End Appearance Front Door (#6250), use the High-End Appearance Side Covers (#6238) to make the rack appear as though it is a high-end server (but in a 19-inch rack format instead of a 24-inch rack).

- The #ERG7 provides an attractive black full height rack door. The door is steel, with a perforated flat front surface. The perforation pattern extends from the bottom to the top of the door to enhance ventilation and provide some visibility into the rack. The non-acoustic door has a depth of about 134 mm (5.3 in).

Rear Door Heat Exchanger

To lead away more heat, a special door named the Rear Door Heat Exchanger (#6858) is available. This door replaces the standard rear door on the rack. Copper tubes that are attached to the rear door circulate chilled water, provided by the customer. The chilled water removes heat from the exhaust air being blown through the servers and attachments mounted in the rack. With industry standard quick couplings, the water lines in the door attach to the customer-supplied secondary water loop.

For details about planning for the installation of the IBM Rear Door Heat Exchanger, see the following website:

http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/iphad_p5/iphadexchangeroverview.html

1.8.3 IBM 42U Slim Rack 7965-94Y

The 2.0-meter (79-inch) model 7965-94Y is compatible with past and present IBM Power Systems servers and provides an excellent 19-inch rack enclosure for your data center. Its 600 mm (23.6 in.) width combined with its 1100 mm (43.3 in.) depth plus its 42 EIA enclosure capacity provides great footprint efficiency for your systems and allows it to be easily located on standard 24-inch floor tiles.

The IBM 42U Slim Rack has a lockable perforated front steel door, providing ventilation, physical security, and visibility of indicator lights in the installed equipment within. In the rear, either a lockable perforated rear steel door (#EC02) or a lockable Rear Door Heat Exchanger (RDHX)(1164-95X) is used. Lockable optional side panels (#EC03) increase the rack's aesthetics, help control airflow through the rack, and provide physical security. Multiple 42U Slim Racks can be bolted together to create a rack suite (indicate #EC04).

Up to six optional 1U PDUs can be placed vertically in the sides of the rack. Additional PDUs can be located horizontally, but they each will use 1U of space in this position.

1.8.4 Feature code 0551 rack

The 1.8 Meter Rack (#0551) is a 36 EIA unit rack. The rack that is delivered as #0551 is the same rack that is delivered when you order the 7014-T00 rack. The included features might vary. Certain features that are delivered as part of the 7014-T00 must be ordered separately with the #0551.

1.8.5 Feature code 0553 rack

The 2.0 Meter Rack (#0553) is a 42 EIA unit rack. The rack that is delivered as #0553 is the same rack that is delivered when you order the 7014-T42 rack. The included features might vary. Certain features that are delivered as part of the 7014-T42 must be ordered separately with the #0553.

1.8.6 Feature code ER05 rack

This feature provides a 19-inch, 2.0 meter high rack with 42 EIA units of total space for installing rack mounted CECs and/or expansion units. The 600mm wide rack fits within a data center's 24" floor tiles and provides better thermal and cable management capabilities. The following features are required on the #ER05:

- #EC01 front door
- #EC02 rear door or #EC05 Rear Door Heat Exchanger (RDHX) indicator

PDUs on the rack are optional. Each #7196 and #7189 PDU consumes one of six vertical mounting bays. Each PDU beyond four will consume 1U of rack space.

If ordering Power Systems equipment in an MES order, use the equivalent rack feature ER05 instead of 7965-94Y so IBM Manufacturing can ship the hardware in the rack.

1.8.7 The AC power distribution unit and rack content

For rack models T00 and T42, 12-outlet PDUs are available. These include the AC power distribution units #9188 and #7188 and the AC Intelligent PDU+ #5889 and #7109.

The Intelligent PDU+ (#5889 and #7109) is identical to #9188 and #7188 PDUs but are equipped with one Ethernet port, one console serial port, and one RS232 serial port for power monitoring.

The PDUs have 12 client-usable IEC 320-C13 outlets. There are six groups of two outlets fed by six circuit breakers. Each outlet is rated up to 10 amps, but each group of two outlets is fed from one 15 amp circuit breaker.

Four PDUs can be mounted vertically in the back of the T00 and T42 racks. Figure 1-5 shows placement of the four vertically mounted PDUs. In the rear of the rack, two additional PDUs can be installed horizontally in the T00 rack and three in the T42 rack. The four vertical mounting locations will be filled first in the T00 and T42 racks. Mounting PDUs horizontally consumes 1U per PDU and reduces the space available for other racked components. When mounting PDUs horizontally, the best approach is to use fillers in the EIA units that are occupied by these PDUs to facilitate proper air-flow and ventilation in the rack.

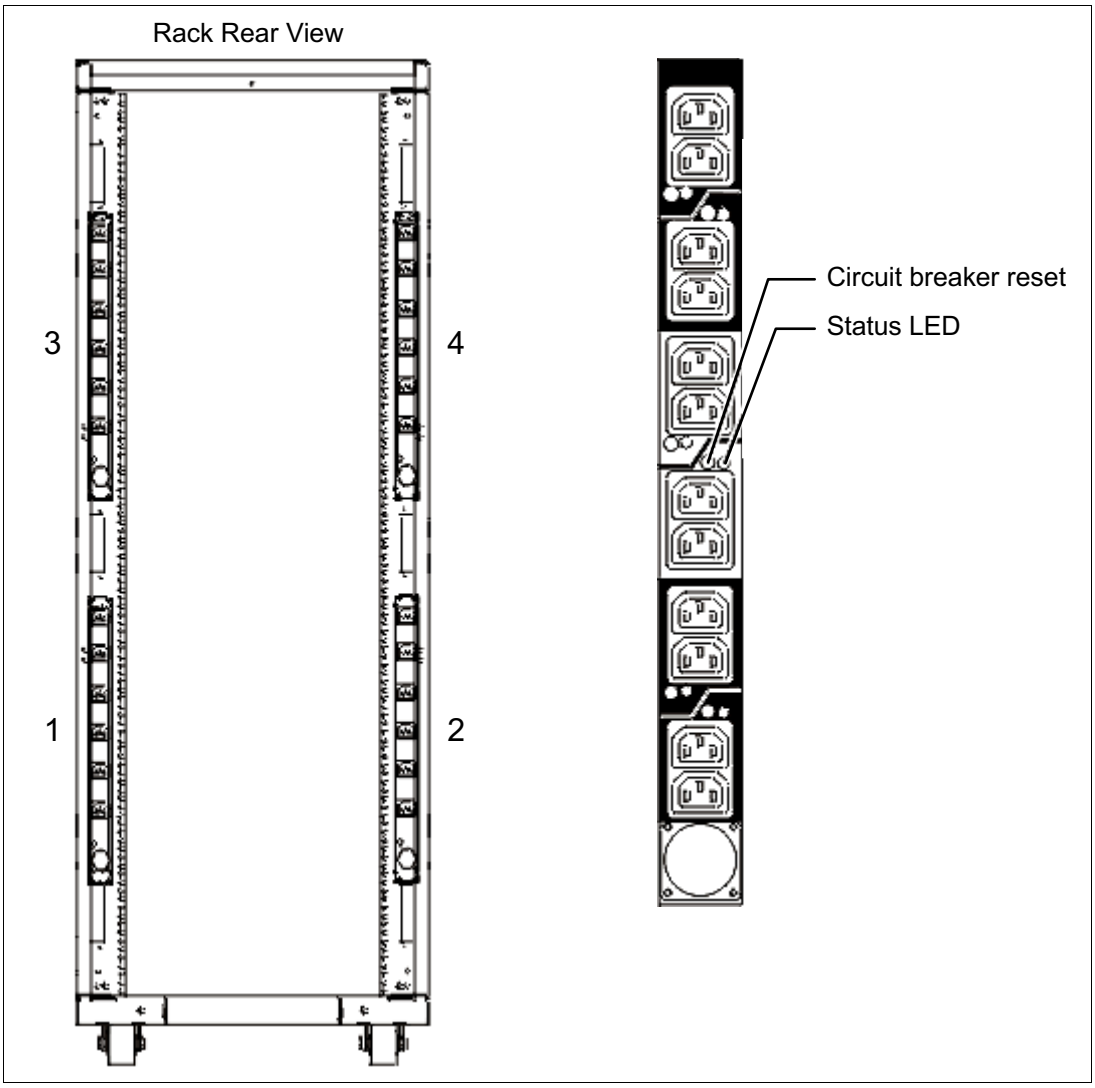


Figure 1-5 PDU placement and PDU view

The PDU receives power through a UTG0247 power-line connector. Each PDU requires one PDU-to-wall power cord. Various power cord features are available for various countries and applications by varying the PDU-to-wall power cord, which must be ordered separately. Each power cord provides the unique design characteristics for the specific power requirements. To match new power requirements and save previous investments, these power cords can be requested with an initial order of the rack or with a later upgrade of the rack features.

Table 1-8 shows the available wall power cord options for the PDU and iPDU features, which must be ordered separately.

Table 1-8 Wall power cord options for the PDU and iPDU features

Feature code	Wall plug	Rated voltage (Vac)	Phase	Rated amperage	Geography
6653	IEC 309, 3P+N+G, 16A	230	3	16 Amps/Phase	Internationally available
6489	IEC309 3P+N+G, 32A	230	3	32 Amps/Phase	EMEA
6654	NEMA L6-30	200-208, 240	1	24 Amps	US, Canada, LA, Japan
6655	RS 3750DP (watertight)	200-208, 240	1	24 Amps	US, Canada, LA, Japan
6656	IEC 309, P+N+G, 32A	230	1	32 Amps	EMEA
6657	PDL	230-240	1	32 Amps	Australia, New Zealand
6658	Korean plug	220	1	30 Amps	North and South Korea
6492	IEC 309, 2P+G, 60A	200-208, 240	1	48 Amps	US, Canada, LA, Japan
6491	IEC 309, P+N+G, 63A	230	1	63 Amps	EMEA

Notes: Ensure that the appropriate power cord feature is configured to support the power being supplied. Based on the power cord that is used, the PDU can supply from 4.8 kVA to 19.2 kVA. The power of all the drawers plugged into the PDU must not exceed the power cord limitation.

The Universal PDUs are compatible with previous models.

To better enable electrical redundancy, each server has two power supplies that must be connected to separate PDUs, which are not included in the base order.

Redundant power supplies: The second power supply for the Power S822 server is optional and not included in the base order.

For maximum availability, a highly desirable approach is to connect power cords from the same system to two separate PDUs in the rack, and to connect each PDU to independent power sources.

For detailed power requirements and power cord details about the 7014 racks, see the Planning for power section in the IBM Power Systems Hardware Information Center website:

http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/topic/p7had/p7hadrpower.htm

For detailed power requirements and power cord details about the 7965-94Y rack, see the Planning for power section in the IBM Power Systems Hardware Information Center website:

http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/topic/p7had/p7hadkickoff795394x.htm

1.8.8 Rack-mounting rules

Consider the following primary rules when you mount the system into a rack:

- The system is designed to be placed at any location in the rack. For rack stability, start filling a rack from the bottom.
- Any remaining space in the rack can be used to install other systems or peripherals, if the maximum permissible weight of the rack is not exceeded and the installation rules for these devices are followed.
- Before placing the system into the service position, be sure to follow the rack manufacturer's safety instructions regarding rack stability.

1.8.9 Useful rack additions

This section highlights several rack addition solutions for IBM Power Systems rack-based systems.

IBM System Storage 7226 Model 1U3 Multi-Media Enclosure

The IBM System Storage 7226 Model 1U3 Multi-Media Enclosure can accommodate up to two tape drives, two RDX removable disk drive docking stations, or up to four DVD-RAM drives.

The IBM System Storage 7226 Multi-Media Enclosure supports LTO Ultrium and DAT160 Tape technology, DVD-RAM, and RDX removable storage requirements on the following IBM systems:

- IBM POWER6 processor-based systems
- IBM POWER7 processor-based systems
- IBM POWER8 processor-based systems
- IBM POWER processor-based BladeCenters (supports SAS tape drive devices only)
- IBM POWER processor-based PureFlex® Systems (supports Fibre Channel and USB devices only)

The 7226 offers an expansive list of drive feature options as shown in Table 1-9.

Table 1-9 Supported drive features for 7226-1U3

Feature code	Description	Status
5619	DAT160 SAS Tape Drive	Available
EU16	DAT160 USB Tape Drive	Available
1420	DVD-RAM SAS Optical Drive	Available
1422	DVD-RAM Slim SAS Optical Drive	Available
5762	DVD-RAM USB Optical Drive	Available
5763	DVD Front USB Port Sled with DVD-RAM USB Drive	Available
5757	DVD RAM Slim USB Optical Drive	Available
8248	LTO Ultrium 5 Half High Fibre Tape Drive	Available
8241	LTO Ultrium 5 Half High SAS Tape Drive	Available
8348	LTO Ultrium 6 Half High Fibre Tape Drive	Available
8341	LTO Ultrium 6 Half High SAS Tape Drive	Available

Feature code	Description	Status
EU03	RDX 3.0 Removable Disk Docking Station	Available

Option descriptions are as follows:

- DAT160 160 GB Tape Drives: With SAS or USB interface options and a data transfer rate up to 12 MBps (assumes 2:1 compression), the DAT160 drive is read/write compatible with DAT160, and DDS4 data cartridges.
- LTO Ultrium 5 Half-High 1.5 TB SAS and FC Tape Drive: With a data transfer rate up to 280 MB/s (assuming a 2:1 compression), the LTO Ultrium 5 drive is read/write compatible with LTO Ultrium 5 and 4 data cartridges, and read-only compatible with Ultrium 3 data cartridges. By using data compression, an LTO-5 cartridge is capable of storing up to 3 TB of data.
- LTO Ultrium 6 Half-High 2.5 TB SAS and FC Tape Drive: With a data transfer rate up to 320 MB/s (assuming a 2.5:1 compression), the LTO Ultrium 6 drive is read/write compatible with LTO Ultrium 6 and 5 media, and read-only compatibility with LTO Ultrium 4. By using data compression, an LTO-6 cartridge is capable of storing up to 6.25 TB of data.
- DVD-RAM: The 9.4 GB SAS Slim Optical Drive with an SAS and USB interface option is compatible with most standard DVD disks.
- RDX removable disk drives: The RDX USB docking station is compatible with most RDX removable disk drive cartridges when used in the same operating system. The 7226 offers the following RDX removable drive capacity options:
  - 320 GB (#EU08)
  - 500 GB (#1107)
  - 1.0 TB (#EU01)
  - 1.5 TB (#EU15)

Removable RDX drives are in a rugged cartridge that inserts in an RDX removable (USB) disk docking station (#1103 or #EU03). RDX drives are compatible with docking stations, installed internally in IBM POWER6, POWER6+, POWER7, POWER7+, and POWER8 servers, where applicable.

Media used in the 7226 DAT160 SAS and USB tape drive features are compatible with DAT160 tape drives installed internally in IBM POWER6, POWER6+, POWER7, POWER7+, and POWER8 servers, and in IBM BladeCenter systems.

Media used in LTO Ultrium 5 Half-High 1.5 TB tape drives are compatible with Half High LTO5 tape drives installed in the IBM TS2250 and TS2350 external tape drives, IBM LTO5 tape libraries, and half-high LTO5 tape drives installed internally in IBM POWER6, POWER6+, POWER7, POWER7+, and POWER8 servers.

Figure 1-6 shows the 7226 Multi-Media Enclosure.

Figure 1-6 7226 Multi-Media Enclosure

The 7226 enclosure offers customer-replaceable unit (CRU) maintenance service to help make installation or replacement of new drives efficient. Other 7226 components are also designed for CRU maintenance.

The IBM System Storage 7226 Multi-Media Enclosure is compatible with most IBM POWER6, POWER6+, POWER7, POWER7+, and POWER8 systems, and also with the IBM BladeCenter models (PS700, PS701, PS702, PS703, and PS704) that offer current level AIX, IBM i, and Linux operating systems.

For a complete list of host software versions and release levels that support the 7226, see the following System Storage Interoperation Center (SSIC) website:

http://www.ibm.com/systems/support/storage/config/ssic/index.jsp

Note: Any of the existing 7216-1U2, 7216-1U3 and 7214-1U2 multimedia drawer are also supported.

Flat panel display options

The IBM 7316 Model TF4 is a rack-mountable flat panel console kit which can also be configured with the tray pulled forward and the monitor folded up, providing full viewing and keying capability for the HMC operator.

The Model TF4 is a follow on product to the Model TF3 and offers the following features:

- Slim, sleek, lightweight monitor design that occupies only 1U (1.75 inches) in a 19-inch standard rack
- A 18.5-inch (409.8 mm x 230.4 mm), flat panel TFT monitor with truly accurate images and virtually no distortion
- The ability to mount the IBM Travel Keyboard in the 7316-TF4 rack keyboard tray
- Support for the following IBM Keyboard/Video/Mouse (KVM) switch:
  - IBM 1x8 Rack Console Switch (#4283)

    Feature 4283, is a 1x8 Console Switch that fits in the 1U space behind the TF4. It is a CAT5 based switch containing eight rack interface (ARI) ports for connecting either PS/2 or USB console switch cables. It supports chaining of servers using IBM Conversion Options switch cable feature 4269. This feature provides four cables that connect a KVM (Keyboard/Video/Mouse) switch to a system, or can be used in a daisy-chain scenario to

connect up to 128 systems to a single KVM switch. It also supports server-side USB attachments.

1.8.10 OEM rack

The system can be installed in a suitable OEM rack, provided that the rack conforms to the EIA-310-D standard for 19-inch racks. This standard is published by the Electrical Industries Alliance. For detailed information, see the IBM Power Systems Hardware Information Center at the following website:

http://publib.boulder.ibm.com/infocenter/systems/scope/hw/index.jsp

The website mentions the following key points:

- The front rack opening must be 451 mm wide ± 0.75 mm (17.75 in. ± 0.03 in.), and the rail-mounting holes must be 465 mm ± 0.8 mm (18.3 in. ± 0.03 in.) apart on-center (horizontal width between the vertical columns of holes on the two front-mounting flanges and on the two rear-mounting flanges). Figure 1-7 on page 21 is a top view showing the specification dimensions.

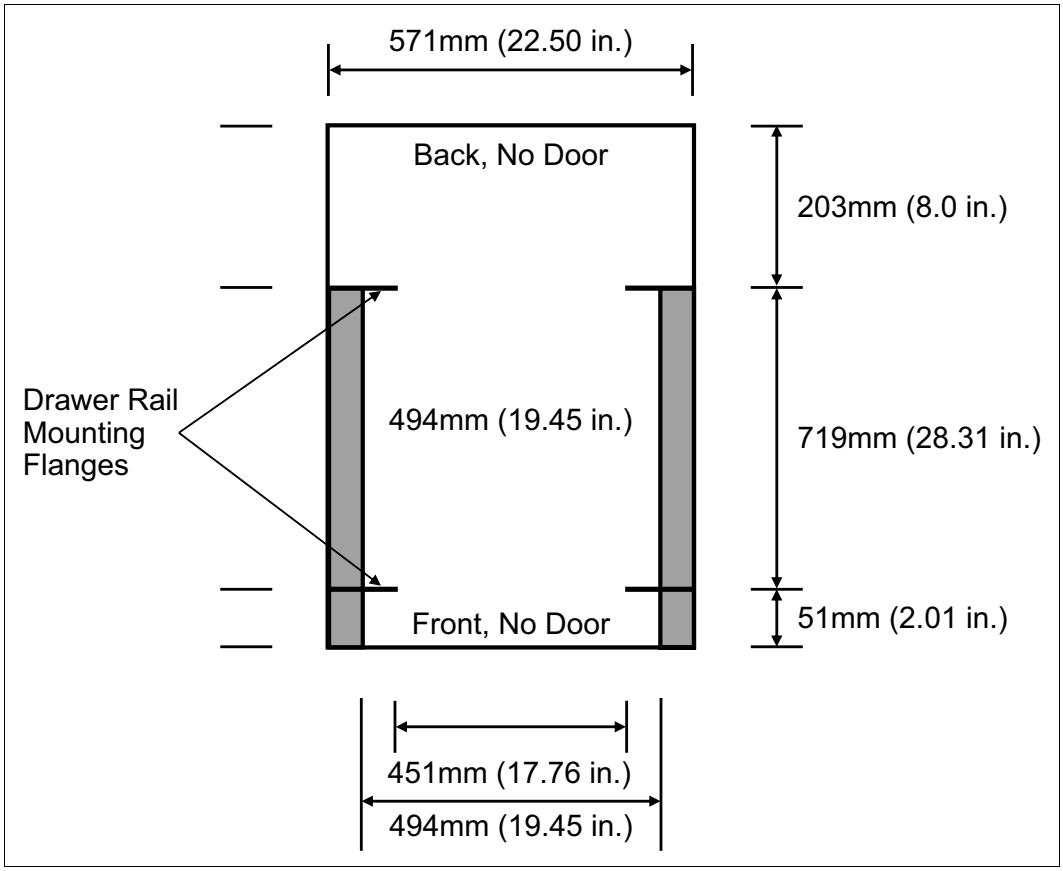


Figure 1-7 Top view of rack specification dimensions (not specific to IBM)

- The vertical distance between the mounting holes must consist of sets of three holes spaced (from bottom to top) 15.9 mm (0.625 in.), 15.9 mm (0.625 in.), and 12.67 mm (0.5 in.) on-center, making each three-hole set of vertical hole spacing 44.45 mm (1.75 in.) apart on center. Rail-mounting holes must be 7.1 mm ± 0.1 mm (0.28 in. ± 0.004 in.) in diameter. Figure 1-8 shows the top front specification dimensions.

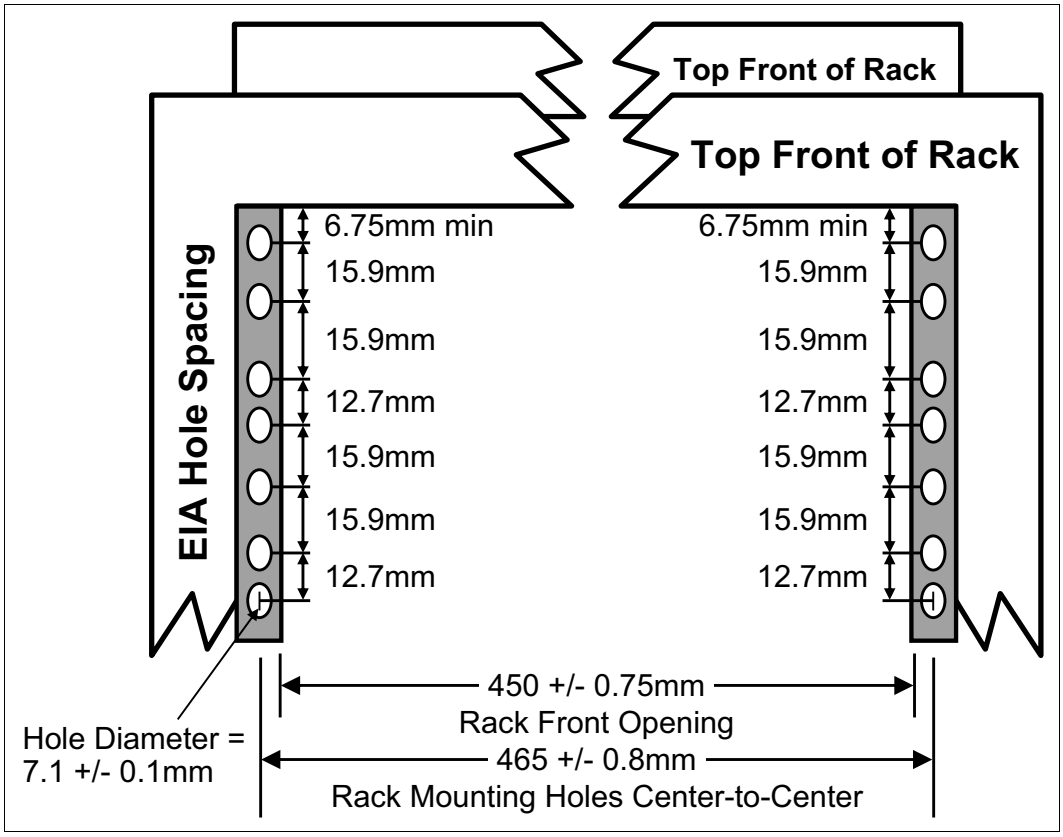


Figure 1-8 Rack specification dimensions, top front view

2

Architecture and technical overview

This chapter discusses the overall system architecture for the Power S822. The bandwidths that are provided throughout the section are theoretical maximums used for reference.

The speeds shown are at an individual component level. Multiple components and application implementation are key to achieving the best performance.

Always do the performance sizing at the application workload environment level and evaluate performance using real-world performance measurements and production workloads.

© Copyright IBM Corp. 2014. All rights reserved.

Figure 2-1 on page 24 shows the logical system diagram for the one socket Power S822.

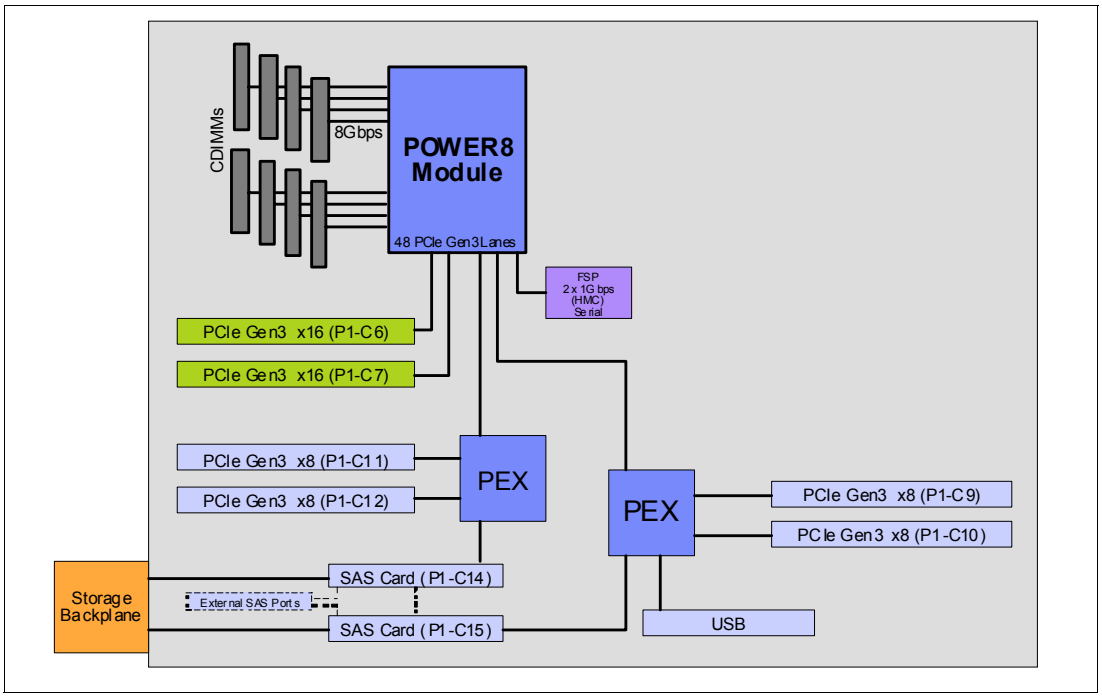


Figure 2-1 Logical system diagram for a one socket Power S822

Figure 2-2 shows the logical system diagram for the two socket Power S822.

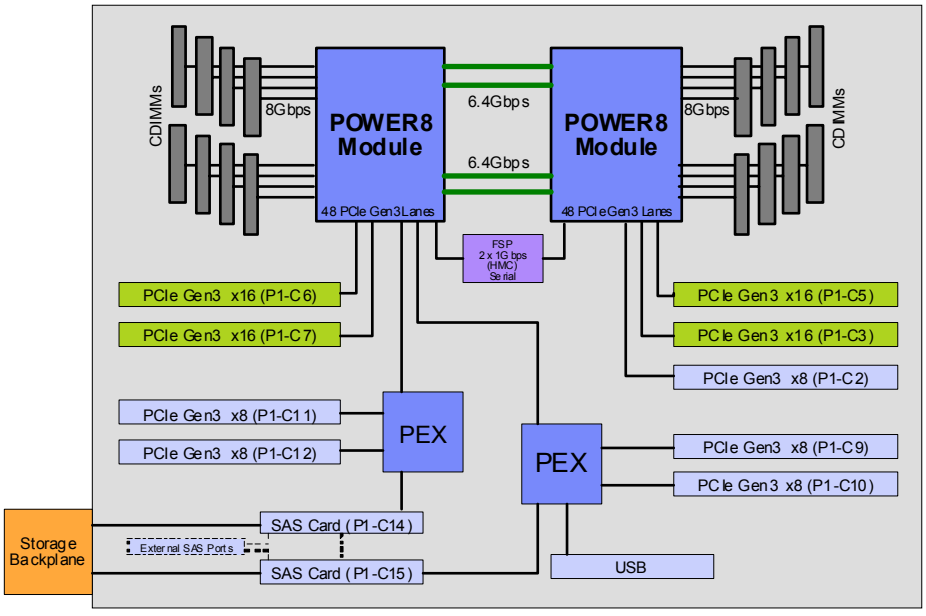


Figure 2-2 Logical system diagram for a two socket Power S822

2.1 The IBM POWER8 processor

This section introduces the latest processor in the IBM Power Systems product family, and discuss its main characteristics and features.

2.1.1 POWER8 processor overview

The POWER8 processor is manufactured using the IBM 22 nm Silicon-On-Insulator (SOI) technology. Each chip is 649 mm^2 and contains 4.2 billion transistors. As shown in Figure 2-3, the chip contains twelve cores, two memory controllers, PCIe Gen3 I/O controllers and an interconnection system that connects all components within the chip. Each core has 512 KB of L2 cache, and all cores share 96 MB of L3 embedded DRAM (eDRAM). The interconnect also extends through module and board technology to other POWER8 processors in addition to DDR3 memory and various I/O devices.

POWER8 systems utilize memory buffer chips to interface between the POWER8 processor and DDR3 or DDR4 memory[1]. Each buffer chip also includes an L4 cache to reduce the latency of local memory accesses.

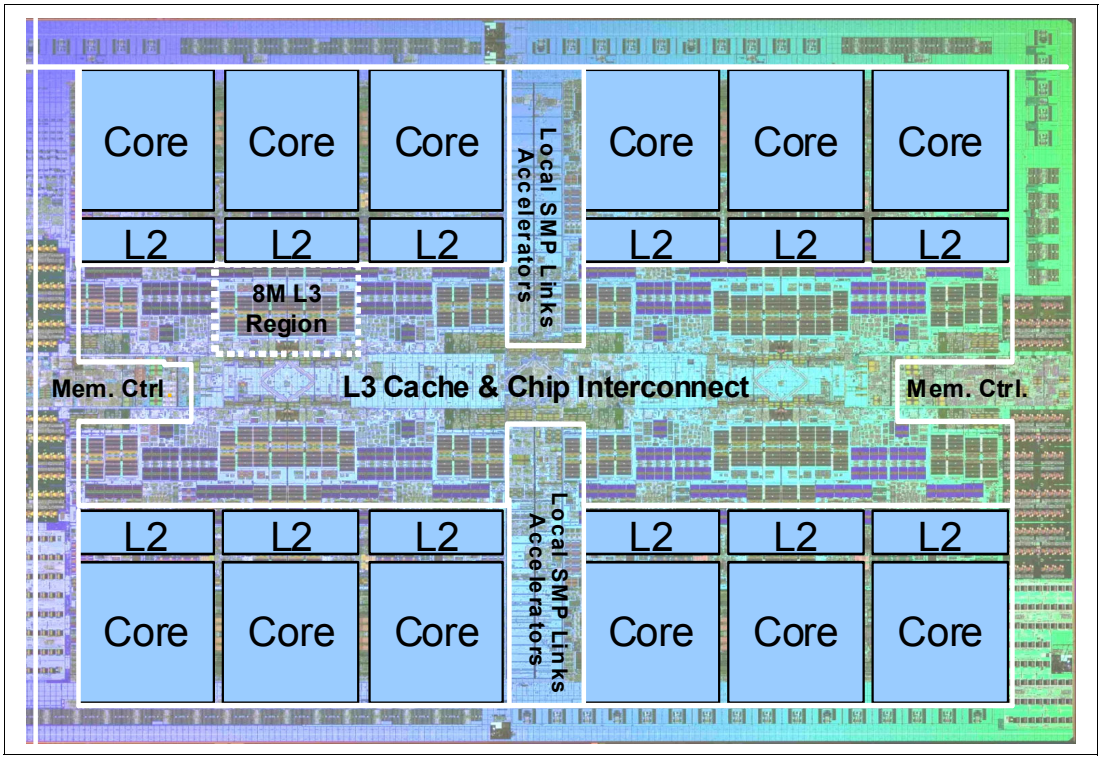


Figure 2-3 The POWER8 processor chip

The POWER8 processor is designed for system offerings from single-socket servers to multi-socket Enterprise servers. It incorporates a triple-scope broadcast coherence protocol over local and global SMP links to provide superior scaling attributes. Multiple-scope coherence protocols reduce the amount of SMP link bandwidth that is required by attempting operations on a limited scope (single chip or multi-chip group) when possible. If the operation cannot complete coherently, the operation is re-issued using a larger scope to complete the operation.

Additional features that can augment performance of the POWER8 processor are:

- Support for DDR3 and DDR4 memory through memory buffer chips that offload the memory support from the POWER8 memory controller

[1] At the time of the publication, the available POWER8 processor-based systems use DDR3 memory.

- L4 cache within the memory buffer chip that reduces the memory latency for local access to memory behind the buffer chip; the operation of the L4 cache is transparent to applications running on the POWER8 processor. Up to 128 MB of L4 cache can be available for each POWER8 processor.
- Hardware transactional memory
- On-chip accelerators, including on-chip encryption, compression, and random number generation accelerators
- Coherent Accelerator Processor Interface, which allow accelerators plugged into a PCIe slot to access the processor bus using a low latency, high speed protocol interface.
- Adaptive power management

As in with the POWER7 processor, the POWER8 processor has a distributed Recovery Unit (RU) that is responsible for saving each processor state and providing a checkpoint for recovery should an error occur.

There are two versions of the POWER8 processor chip. Both chips use the same building blocks. The scale-out systems use a 6-core version of POWER8. The 6-core chip is installed in pairs in a Dual Chip Module (DCM) that plugs into a socket in the planar of the systems. Functionally, it works as a single chip.

Figure 2-4 shows a graphic representation of the 6-core processor.

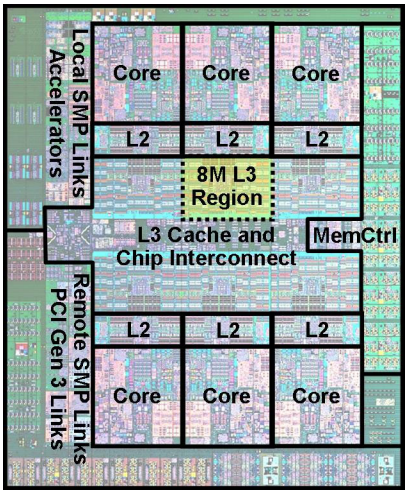


Figure 2-4 6-core POWER8 processor chip

Table 2-1 summarizes the technology characteristics of the POWER8 processor.

Table 2-1 Summary of POWER8 processor technology

Technology	POWER8 processor
Die size	649 mm^2
Fabrication technology	► 22 nm lithography ► Copper interconnect ► Silicon-on-Insulator ► eDRAM
Maximum processor cores	6 or 12
Maximum execution threads core/chip	8/96
Maximum L2 cache core/chip	512 KB/6 MB
Maximum On-chip L3 cache core/chip	8 MB/96 MB
Maximum L4 cache per chip	128 MB
Maximum memory controllers	2
SMP design-point	16 sockets with IBM POWER8 processors
Compatibility	With prior generation of POWER processor

2.1.2 POWER8 processor core

The POWER8 processor core is a 64-bit implementation of the IBM Power Instruction Set Architecture (ISA) Version 2.07 and has the following features:

- Multi-threaded design, capable of up to eight-way simultaneous multithreading (SMT)
- 32 KB, eight-way set-associative L1 instruction cache
- 64 KB, eight-way set-associative L1 data cache
- Enhanced prefetch, with instruction speculation awareness and data prefetch depth awareness
- Enhanced branch prediction, using both local and global prediction tables with a selector table to choose the best predictor
- Improved out-of-order execution
- Two symmetric fixed-point execution units
- Two symmetric load/store units and two load units, all four of which can also run simple fixed-point instructions
- An integrated, multi-pipeline vector-scalar floating point unit for running both scalar and SIMD-type instructions, including the Vector Multimedia eXtension (VMX) instruction set and the improved Vector Scalar eXtension (VSX) instruction set, and capable of up to eight floating point operations per cycle (four double precision or eight single precision)
- In-core Advanced Encryption Standard (AES) encryption capability
- Hardware data prefetching with 16 independent data streams and software control
- Hardware decimal floating point (DFP) capability.

More information about Power ISA Version 2.07 can be found at:

https://www.power.org/documentation/power-isa-version-2-07/

Figure 2-8 on page 32 shows a picture of the POWER8 core, with some of the functional units highlighted.

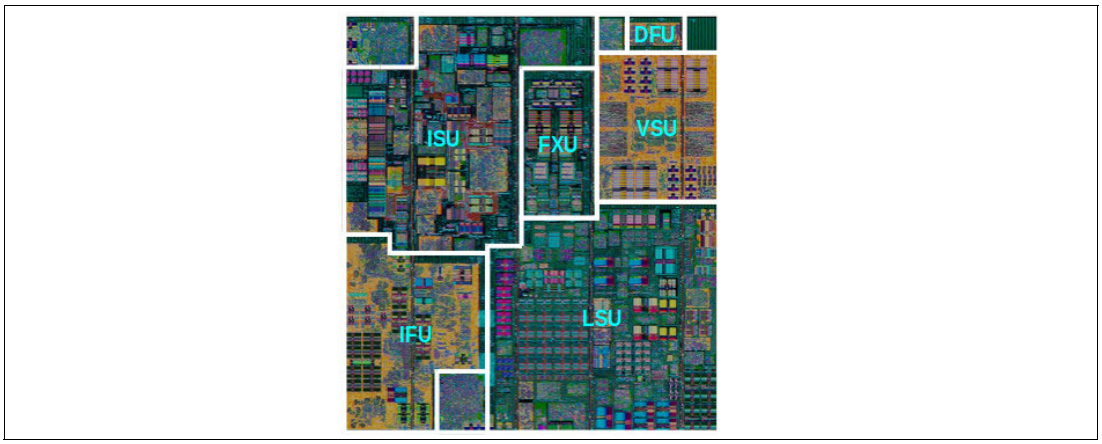


Figure 2-5 POWER8 processor core

2.1.3 Simultaneous multithreading

POWER8 processor advancements in multi-core and multi-thread scaling are remarkable. A significant performance opportunity comes from parallelizing workloads to enable the full potential of the microprocessor, as well as the large memory bandwidth. Application scaling is influenced by both multi-core and multi-thread technology.

Simultaneous Multithreading (SMT) allows a single physical processor core to simultaneously dispatch instructions from more than one hardware thread context. With SMT, each POWER8 core can present eight hardware threads. Because there are multiple hardware threads per physical processor core, additional instructions can run at the same time. SMT is primarily beneficial in commercial environments where the speed of an individual transaction is not as critical as the total number of transactions performed. SMT typically increases the throughput of workloads with large or frequently changing working sets, such as database servers and web servers.

Table 2-2 shows a comparison between the different POWER processors in terms of SMT capabilities supported by each processor architecture.

Table 2-2 SMT levels supported by POWER processors

Technology	Cores/system	Maximum SMT mode	Maximum hardware threads per partition
IBM POWER4	32	Single Thread (ST)	32
IBM POWER5	64	SMT2	128
IBM POWER6	64	SMT2	128
IBM POWER7	256	SMT4	1024
IBM POWER8	192	SMT8	1536

The architecture of the POWER8 processor, with its larger caches, larger cache bandwidth and faster memory allows threads to have faster access to memory resources, that translates in a more efficient use of threads. Because of that, POWER8 allows more threads per core to run concurrently, increasing the total throughput of the processor and of the system.

2.1.4 Memory access

On the Power S822, each POWER8 module has two memory controllers, each connected to four memory channels. Each memory channel operates at 1600 MHz and connects to a DIMM. Each DIMM on a POWER8 system has a memory buffer that is responsible for many functions that were previously on the memory controller, such as scheduling logic and energy management. The memory buffer also has 16 MB of L4 cache.

At the time of writing each memory channel can address up to 64 GB. Therefore the Power S822 is capable of addressing up to 1 TB of total memory.

Figure 2-6 gives a simple overview of the POWER8 processor memory access structure in the Power S822 server.

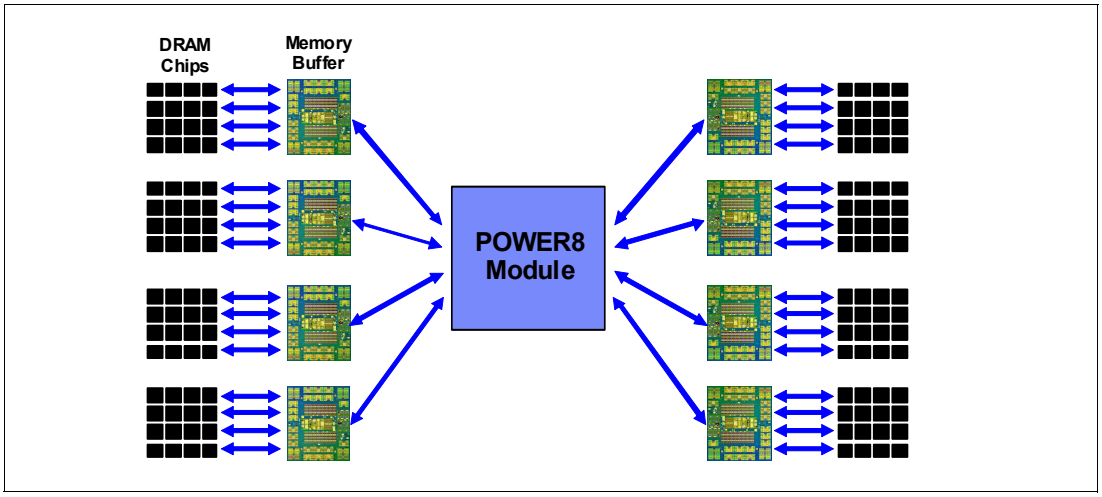


Figure 2-6 Overview of POWER8 memory access structure

2.1.5 On-chip L3 cache innovation and Intelligent Cache

Similar to POWER7 and POWER7+, the POWER8 processor utilizes a breakthrough in material engineering and microprocessor fabrication to implement the L3 cache in eDRAM and place it on the processor die. L3 cache is critical to a balanced design, as is the ability to provide good signaling between the L3 cache and other elements of the hierarchy, such as the L2 cache or SMP interconnect.

The on-chip L3 cache is organized into separate areas with differing latency characteristics. Each processor core is associated with a fast 8 MB local region of L3 cache (FLR-L3) but also has access to other L3 cache regions as shared L3 cache. Additionally, each core can negotiate to use the FLR-L3 cache associated with another core, depending on reference patterns. Data can also be cloned to be stored in more than one core's FLR-L3 cache, again depending on reference patterns. This Intelligent Cache management enables the POWER8 processor to optimize the access to L3 cache lines and minimize overall cache latencies.

Figure 2-3 on page 25 and Figure 2-4 on page 26 show the on-chip L3 cache, and highlight one fast 8 MB L3 region closest to a processor core.

The innovation of using eDRAM on the POWER8 processor die is significant for several reasons:

- Latency improvement

  A six-to-one latency improvement occurs by moving the L3 cache on-chip compared to L3 accesses on an external (on-ceramic) ASIC.

- Bandwidth improvement

  A 2x bandwidth improvement occurs with on-chip interconnect. Frequency and bus sizes are increased to and from each core.

- No off-chip driver or receivers

  Removing drivers or receivers from the L3 access path lowers interface requirements, conserves energy, and lowers latency.

- Small physical footprint

  The performance of eDRAM when implemented on-chip is similar to conventional SRAM but requires far less physical space. IBM on-chip eDRAM uses only a third of the components that conventional SRAM uses, which has a minimum of six transistors to implement a 1-bit memory cell.

- Low energy consumption

  The on-chip eDRAM uses only 20% of the standby power of SRAM.

2.1.6 L4 cache and memory buffer

POWER8 processor-based systems introduce an additional level in memory hierarchy. The L4 cache is implemented together with the memory buffer in the Custom DIMM (CDIMM). Each memory buffer contains 16 MB of L4 cache, and on a Power S822 you can have up to 128 MB of L4 cache.

Figure 2-7 shows a picture of the memory buffer, where you can see the 16 MB L4 cache, as well as processor links and memory interfaces.

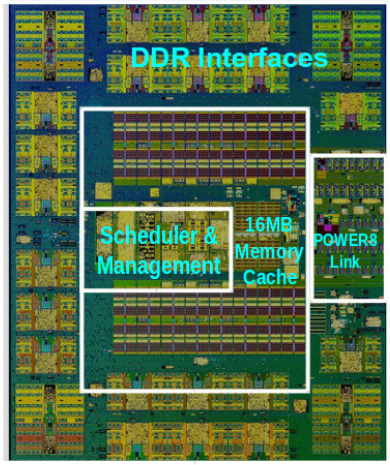


Figure 2-7 Memory buffer chip

Table 2-3 on page 31 shows a comparison of the different levels of cache in the POWER7, POWER7+ and POWER8 processors.

Table 2-3 POWER8 cache hierarchy

Cache	POWER7	POWER7+	POWER8
L1 instruction cache: Capacity/associativity	32 KB, 4-way	32 KB, 4-way	32 KB, 8-way
L1 data cache: Capacity/associativity bandwidth	32 KB, 8-way 2 16 B reads or 1 16 B writes per cycle	32 KB, 8-way 2 16 B reads or 1 16 B writes per cycle	64 KB, 8-way 4 16 B reads or 1 16 B writes per cycle
L2 cache: Capacity/associativity bandwidth	256 KB, 8-way Private 32 B reads and 16 B writes per cycle	256 KB, 8-way Private 32 B reads and 16 B writes per cycle	512 KB, 8-way Private 64 B reads and 16 B writes per cycle
L3 cache: Capacity/associativity bandwidth	On-Chip 4 MB/core, 8-way 16 B reads and 16 B writes per cycle	On-Chip 10 MB/core, 8-way 16 B reads and 16 B writes per cycle	On-Chip 8 MB/core, 8-way 32 B reads and 32 B writes per cycle
L4 cache: Capacity/associativity bandwidth	N/A	N/A	On-Chip 16 MB/buffer chip, 16-way Up to 8 buffer chips per socket

For more information on the POWER8 memory subsystem, see 2.2, “Memory subsystem” on page 34.

2.1.7 Hardware transactional memory

Transactional memory is an alternative to lock-based synchronization. It attempts to simplify parallel programming by grouping read and write operations and running them like a single operation. Transactional memory is like database transactions where all shared memory accesses and their effects are either committed all together or discarded as a group. All threads can enter the critical region simultaneously. If there are conflicts in accessing the shared memory data, threads try accessing the shared memory data again or are stopped without updating the shared memory data. Therefore, transactional memory is also called a lock-free synchronization. Transactional memory can be a competitive alternative to lock-based synchronization.

Transactional memory provides a programming model that makes parallel programming easier. A programmer delimits regions of code that access shared data and the hardware executes these regions atomically and in isolation, buffering the results of individual instructions, and retrying execution if isolation is violated. Generally, transactional memory allows programs to use a programming style that is close to coarse-grained locking to achieve performance that is close to fine-grained locking.

Most implementations of transactional memory are based on software. The POWER8 processor-based systems provide a hardware-based implementation of transactional memory, that is more efficient than the software implementations and requires no interaction with the processor core, therefore allowing the system to operate in maximum performance.

2.1.8 Coherent Accelerator Processor Interface

The Coherent Accelerator Interface Architecture (CAIA) defines a coherent accelerator interface structure for attaching special to the POWER systems.

The Coherent Accelerator Processor Interface (CAPI) provides the ability to attach accelerators that have coherent shared memory access with the processors in the server and share full virtual address translation with these processors, using a standard PCIe Gen3 bus.

Applications can have customized functions in Field Programmable Gate Arrays (FPGA) and be able to enqueue work requests directly in shared memory queues to the FPGA, and using the same effective addresses (pointers) it would use for any of its threads running on a host processor. From the practical perspective, CAPI allows a specialized hardware accelerator to be seen as an additional processor in the system, with access to the main system memory, and coherent communication with other processors in the system.

The benefits of using CAPI include the ability to access shared memory blocks directly from the accelerator, perform memory transfers directly between the accelerator and processor cache, and reduction on the code path length between the adapter and the processors, since the adapter is not operating as a traditional I/O device, and there is no device driver layer to perform processing. It also presents a simpler programming model.

Figure 2-8 shows a high level view on how an accelerator communicates with the POWER8 processor through CAPI. The POWER8 processor provides a Coherent Attached Processor Proxy (CAPP), that is responsible for extending the coherence in the processor communications to an external device. The coherency protocol is tunneled over standard PCIe Gen3, effectively making the accelerator part of the coherency domain.

The accelerator adapter implements the Power Service Layer (PSL), that provides address translation and system memory cache for the accelerator functions. The custom processors on the board, consisting of an FPGA or an Application Specific Integrated Circuit (ASIC) use this layer to access shared memory regions, cache areas as if they were a processor in the system. This ability greatly enhances the performance of the data access for the device and simplifies the programming effort to use the device. Instead of treating the hardware accelerator as an I/O device, it is treated as a processor. That eliminates the requirement of a device driver to perform communication, as well as the need for Direct Memory Access that requires system calls to the operating system kernel. By removing these layers, the data transfer operation requires much less clock cycles in the processor, greatly improving the I/O performance.

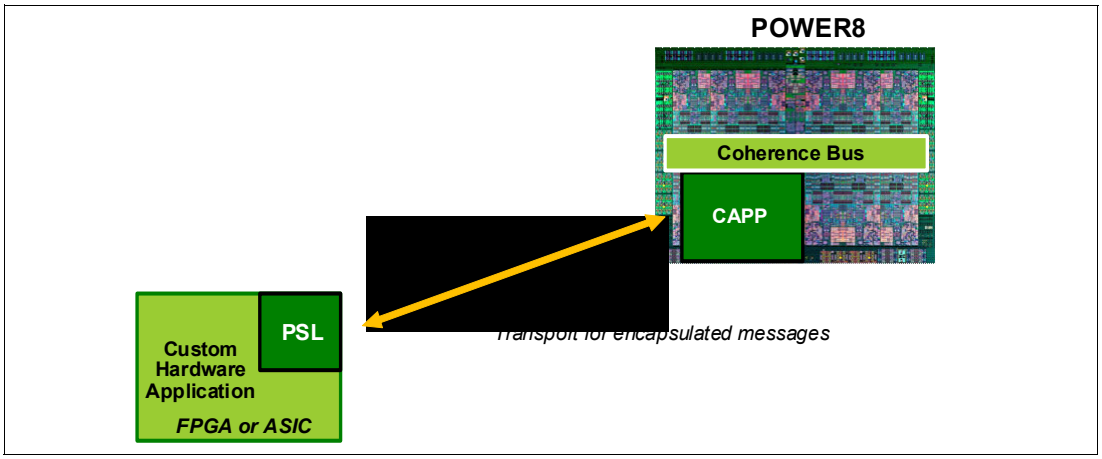


Figure 2-8 CAPI accelerator attached to the POWER8 processor

The implementation of CAPI on the POWER8 processor allows hardware companies to develop solutions for specific application demands and leverage the performance of the POWER8 processor for general applications as well as the custom acceleration of specific

functions using a hardware accelerator, with a simplified programming model and efficient communication with the processor and memory resources.

2.1.9 Power management and system performance

The POWER8 processor has power saving and performance enhancing features that can be used to lower overall energy usage, while yielding higher performance when needed. The following modes can be enabled and modified in order to use these features.

Dynamic Power Saver: Favor Performance

This mode is intended to provide the best performance. If the processor is being used even moderately, the frequency will be raised to the maximum frequency possible to provide the best performance. If the processors are very lightly used, the frequency will be lowered to the minimum frequency, which is potentially far below the nominal shipped frequency, to save energy. Note that the top frequency achieved is based on system type and is affected by environmental conditions. Also note that when running at the maximum frequency, significantly more energy is being consumed, which means this mode can potentially cause an increase in overall energy consumption.

Dynamic Power Saver: Favor Power

This mode is intended to provide the best performance per watt consumed. The processor frequency is adjusted based on the processor utilization to maintain the workload throughput without using more energy than required to do so. At very high processor utilization levels, the frequency will be raised above nominal, just as in the favor performance mode above. Likewise, at very low processor utilization levels, the frequency will be lowered to the minimum frequency. The frequency ranges are the same for the two Dynamic Power Saver modes, but the algorithm that determines which frequency to set is different.

Dynamic Power Saver: Tunable Parameters

The modes just discussed (“Dynamic Power Saver: Favor Performance” and “Dynamic Power Saver: Favor Power”) are tuned to provide both energy savings and performance increases. However, there may be situations where only top performance is of concern, or, conversely, where peak power consumption is an issue. The tunable parameters can be used to modify the setting of the processor frequency in these modes to meet these various objectives. Note that modifying these parameters should be done only by advanced users. We suggest that, if there are issues that need to be addressed by the Tunable Parameters, IBM should be directly involved in the parameter value selection.

Idle Power Saver

This mode is intended to save the maximum amount of energy when the system is nearly completely idle. When the processors are found to be nearly idle, the frequency of all processors is lowered to the minimum. Additionally, workloads are dispatched onto a smaller number of processor cores so that the other processor cores can be put into a low energy usage state. When processor utilization increases, the process is reversed: The processor frequency is raised back up to nominal, and the workloads are spread out once again over all of the processor cores. There is no performance boosting aspect in this mode, but entering or exiting this mode may affect overall performance. The delay times and utilization levels for entering and exiting this mode can be adjusted to allow for more or less aggressive energy savings.

The controls for all modes listed above are available on the Advanced System Management Interface and are described in more detail in a white paper available at:

http://public.dhe.ibm.com/common/ssi/ecm/en/pow03039usen/POW03039USEN.PDF

For more details, see chapter 2.11, “Energy management” on page 70.

2.1.10 Comparison of the POWER8, POWER7+ and POWER7 processors

Table 2-3 on page 31 shows comparable characteristics between the generations of POWER8, POWER7+, and POWER7 processors.

Comparison of technology for the POWER8 processor and the prior generations

Characteristics	POWER8	POWER7+	POWER7
Technology	22 nm	32 nm	45 nm
Die size	649 mm^2	567 mm^2	567 mm^2
Number of transistors	4.2 billion	2.1 billion	1.2 billion
Maximum cores	12	8	8
Maximum SMT threads per core	8 threads	4 threads	4 threads
Maximum frequency	4.15 GHz	4.4 GHz	4.25 GHz
L2 Cache	512 KB per core	256 KB per core	256 KB per core
L3 Cache	8 MB of FLR-L3 cache per core with each core having access to the full 96 MB of L3 cache, on-chip eDRAM	10 MB of FLR-L3 cache per core with each core having access to the full 80 MB of L3 cache, on-chip eDRAM	4 MB or 8 MB of FLR-L3 cache per core with each core having access to the full 32 MB of L3 cache, on-chip eDRAM
Memory support	DDR3 and DDR4	DDR3	DDR3
I/O bus	PCIe Gen3	GX++	GX++

2.2 Memory subsystem

The IBM Power S822 is a two socket system that supports up to two POWER8 processor modules. The server supports a maximum of 16 DDR3 CDIMM slots, with 8 DIMM slots per installed processor. Memory features supported are 16 GB, 32 GB, an 64 GB and run at speeds of 1600 MHz, allowing for a maximum system memory of 1024 GB.

These servers support an optional feature called Active Memory Expansion (#4793) that allows the effective maximum memory capacity to be much larger than the true physical memory. This feature executes innovative compression and decompression of memory content by using a dedicated coprocessor to provide memory expansion up to 125%, depending on the workload type and its memory utilization. As an example, a server with 256 GB RAM physically installed can effectively be expanded over 512 GB RAM. This approach can enhance virtualization and server consolidation by allowing a partition to do significantly more work with the same physical amount of memory or a server to run more partitions and do more work with the same physical amount of memory.

2.2.1 Custom DIMM (CDIMM)

CDIMMs stand for Custom DIMMs and are an innovative memory DIMMs that do not only house industry standard DRAM memory chips but also includes a set of components that allow for higher bandwidth and lower latency communications:

- Memory Scheduler
- Memory Management (RAS Decisions & Energy Management)
- Buffer Cache

By adopting this architecture for the memory DIMMs, several decisions and processes regarding memory optimizations are executed internally into the CDIMM, saving bandwidth and allowing for faster processor to memory communications. This also allows for a more robust RAS. For more information on RAS, refer to Chapter 4, “Continuous availability and manageability” on page 107.

A detailed diagram of the CDIMM available for the Power S822 can be seen in Figure 2-9.

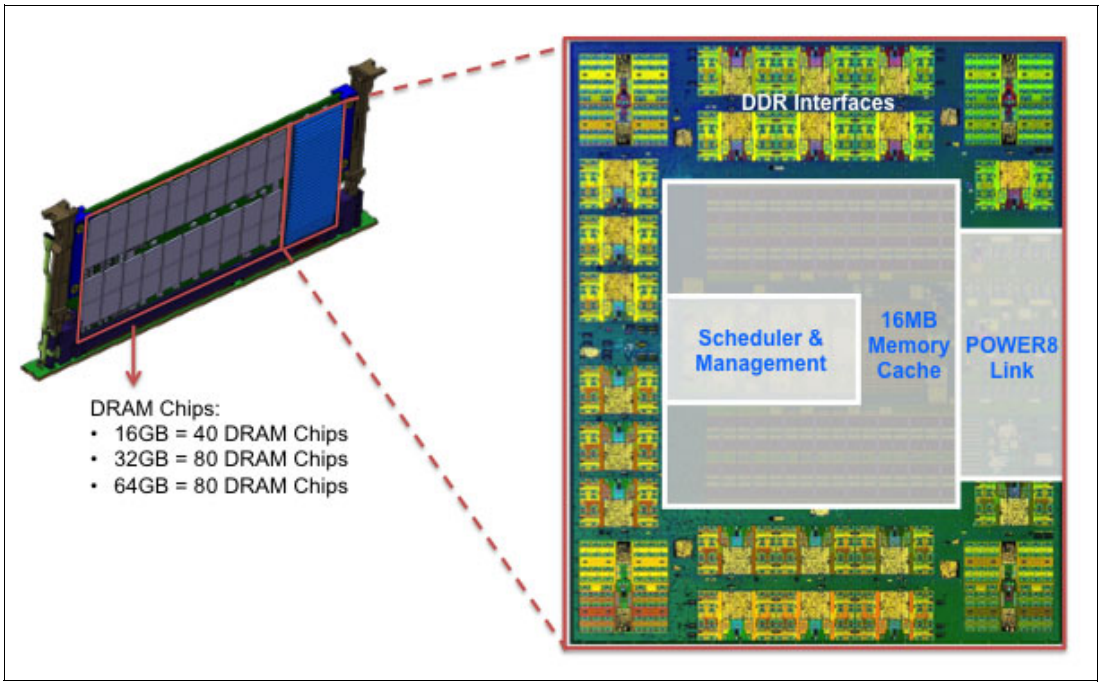


Figure 2-9 Short CDIMM Diagram

The Buffer Cache is a L4 cache and is built on eDRAM technology (same as L3 cache) which has lower latency than regular SRAM. Each CDIMM has 16 MB of L4 cache and a fully populated Power S822 server (2 processor modules and 16 CDIMMs) will have 128 MB of L4 Cache. The L4 Cache performs several functions that have direct impact in performance and brings a series of benefits for the Power S822:

- Reduce energy consumption by reducing the amount of memory requests.
- Increase memory write performance by acting as a cache and by grouping several random writes into larger transactions.
- Partial write operations that target the same cache block will be 'gathered' within the L4 cache before having to be written to memory, becoming a single write operation.
- Reduces latency on memory access. Memory access for cached blocks have up to 55% lower latency than non-cached blocks.

2.2.2 Memory placement rules

The following memory options are orderable:

- 16 GB CDIMM, 1600 MHz DDR3 DRAM (#EM83)
- 32 GB CDIMM, 1600 MHz DDR3 DRAM (#EM84)
- 64 GB CDIMM, 1600 MHz DDR3 DRAM (#EM85)

All memory must be ordered in pairs with a minimum of 32 GB memory required for a Power S822 system.

The supported maximum memory is as follows:

- One processor module installed: 512 GB (eight 64 GB CDIMMs)
- Two processor modules installed: 1024 GB (sixteen 64 GB CDIMMs)

The basic rules for memory placement follows:

- Each feature code equates to a single physical CDIMM.
- All memory features must be ordered in pairs.
- All memory CDIMMs must be installed in pairs.
- Each CDIMM within a pair must be of the same capacity.

Figure 2-10 on page 37 shows the physical memory DIMM topology.

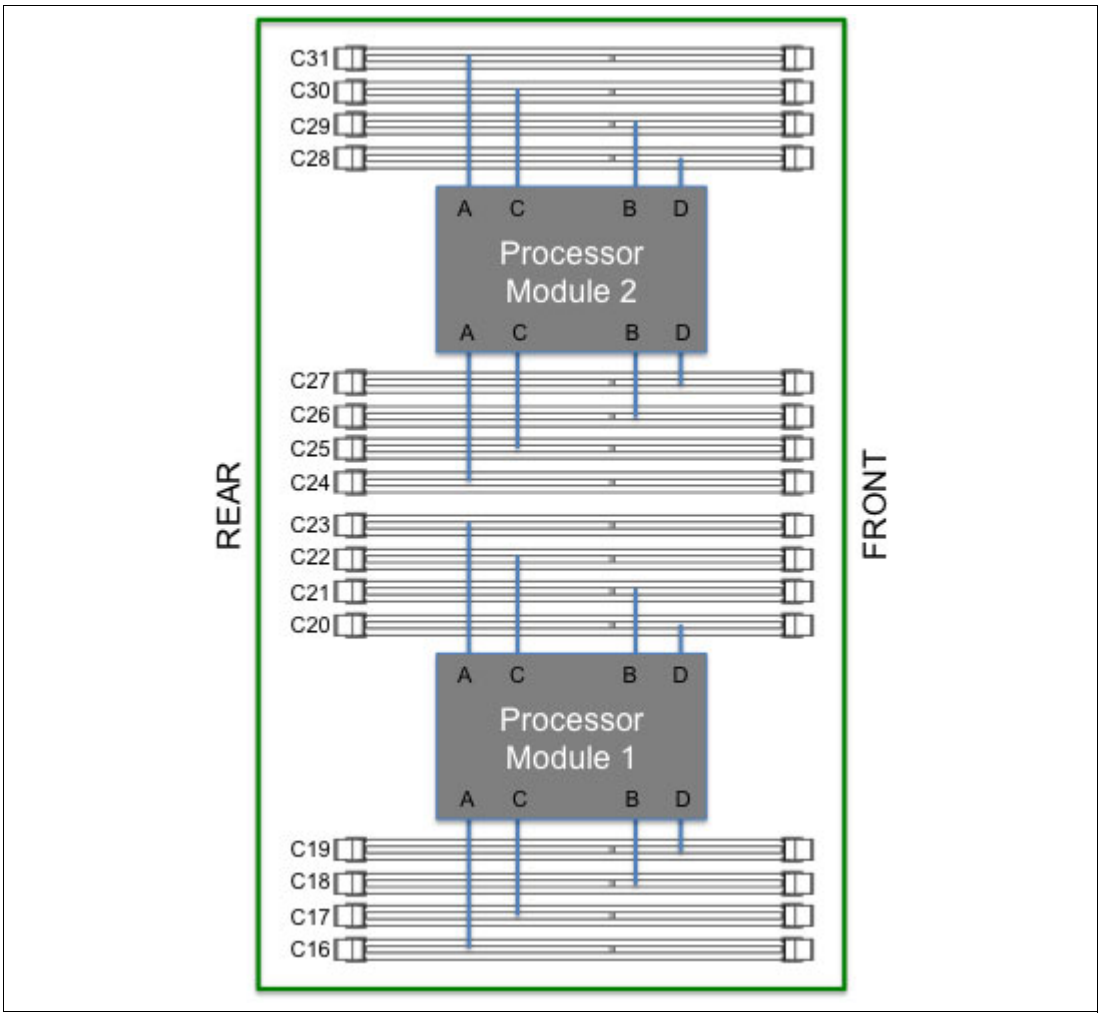


Figure 2-10 Memory DIMM topology for the Power S822

In general, the best approach is to install memory evenly across all processors in the system. Balancing memory across the installed processors allows memory access in a consistent manner and typically results in the best possible performance for your configuration. Account for any plans for future memory upgrades when you decide which memory feature size to use at the time of the initial system order.

For systems with a single processor installed, the plugging order for memory DIMMS is as follows:

- First CDIMM pair is identical and installed at C16 and C18.
- Next CDIMM pair is identical and installed at C21 and C23.
- Next CDIMM pair is identical and installed at C17 and C19.
- Next CDIMM pair is identical and installed at C20 and C22.

For systems with two processors installed, the plugging order for memory DIMMS is as follows:

- First CDIMM pair is identical and installed at C16 and C18.
- Next CDIMM pair is identical and installed at C24 and C26.
- Next CDIMM pair is identical and installed at C21 and C23.
- Next CDIMM pair is identical and installed at C29 and C31.
- Next CDIMM pair is identical and installed at C17 and C19.

- Next CDIMM pair is identical and installed at C25 and C27.
- Next CDIMM pair is identical and installed at C20 and C22.
- Next CDIMM pair is identical and installed at C28 and C30.

2.2.3 Memory bandwidth

The POWER8 processor has exceptional cache, memory, and interconnect bandwidths. Table 2-4 shows the maximum bandwidth estimates for a single core on the Power S822 system.

Table 2-4 Power S822 single core bandwidth estimates

Single Core	Power S822	Power S822
	1 core @ 3.42 GHz	1 core @ 3.89 GHz
L1 (data) cache	164.2 GBps	186.7 GBps
L2 cache	164.2 GBps	186.7 GBps
L3 cache	218.9 GBps	249.0 GBps

The bandwidth figures for the caches are calculated as follows:

- L1 cache: In one clock cycle, two 16-byte load operations and one 16-byte store operation can be accomplished. The value varies depending on the clock of the core and the formula is as follows:

```
3.42 GHz Core: (2 * 16 B + 1 * 16 B) * 3.42 GHz = 164.2 GBps
3.89 GHz Core: (2 * 16 B + 1 * 16 B) * 3.89 GHz = 186.7 GBps
```

- L2 cache: In one clock cycle, one 32-byte load operation and one 16-byte store operation can be accomplished. The value varies depending on the clock of the core and the formula is as follows:

```
3.42 GHz Core: (1 * 32 B + 1 * 16 B) * 3.42 GHz = 164.2 GBps
3.89 GHz Core: (1 * 32 B + 1 * 16 B) * 3.89 GHz = 186.7 GBps
```

- L3 cache: One 32-byte load operation and one 32-byte store operation can be accomplished at each clock cycle and the formula is as follows:

```
3.42 GHz Core: (1 * 32 B + 1 * 32 B) * 3.42 GHz = 218.9 GBps
3.89 GHz Core: (1 * 32 B + 1 * 32 B) * 3.89 GHz = 249.0 GBps
```

For the whole system, considering a Power S822 populated with two processor modules, the overall bandwidths are as follows:

Table 2-5 Power S822 total bandwidth estimates

Total Bandwidths	Power S822	Power S822
	20 cores @ 3.42 GHz	12 cores @ 3.89 GHz
L1 (data) cache	3,284 GBps	2,240 GBps
L2 cache	3,284 GBps	2,240 GBps
L3 cache	4,352 GBps	2,988 GBps
Total Memory	384 GBps	384 GBps
SMP Interconnect	38.4 GBps	38.4 GBps

SMP interconnect: The POWER8 processor has four 2-byte lanes working at 6.4 GHz with one spare lane (total 3 active lanes). The bandwidth formula is calculated as follows:

```
3 lanes * 2 Bytes * 6.4 GHz = 38.4 GBps
```

2.3 System Bus

This section provides more information related to the internal buses.

The Power S822 systems have internal I/O connectivity through Peripheral Component Interconnect Express Gen3 (PCI Express Gen3 or PCIe Gen3) slots and also external connectivity through SAS adapters.

The internal I/O subsystem on the Power S822 is connected to the PCIe Controllers on a POWER8 processor module in the system. Each POWER8 processor module has a bus that has 48 PCIe lanes running at 8 Gbps full-duplex and provides 96 GBps of I/O connectivity to the PCIe slots, SAS internal adapters, and USB ports.

Some PCIe devices are connected directly to the PCIe Gen3 buses on the processors, while other devices are connected to these buses via PCIe Gen3 Switches. The PCIe Gen3 Switches are high speed devices (ranging from 512 GBps to 768 GBps each) that allow for the optimal utilization of the processors PCIe Gen3 x16 buses by grouping slower x8 or x4 devices that would plug into a x16 slot and not use its full bandwidth. For more details on which slots are connected directly to the #and which ones are attached to a PCIe Gen3 Switches (referred as PEX), refer to Chapter 2, “Architecture and technical overview” on page 23.

A diagram comparing the POWER7 and POWER8 I/O buses architecture can be seen on Figure 2-11.

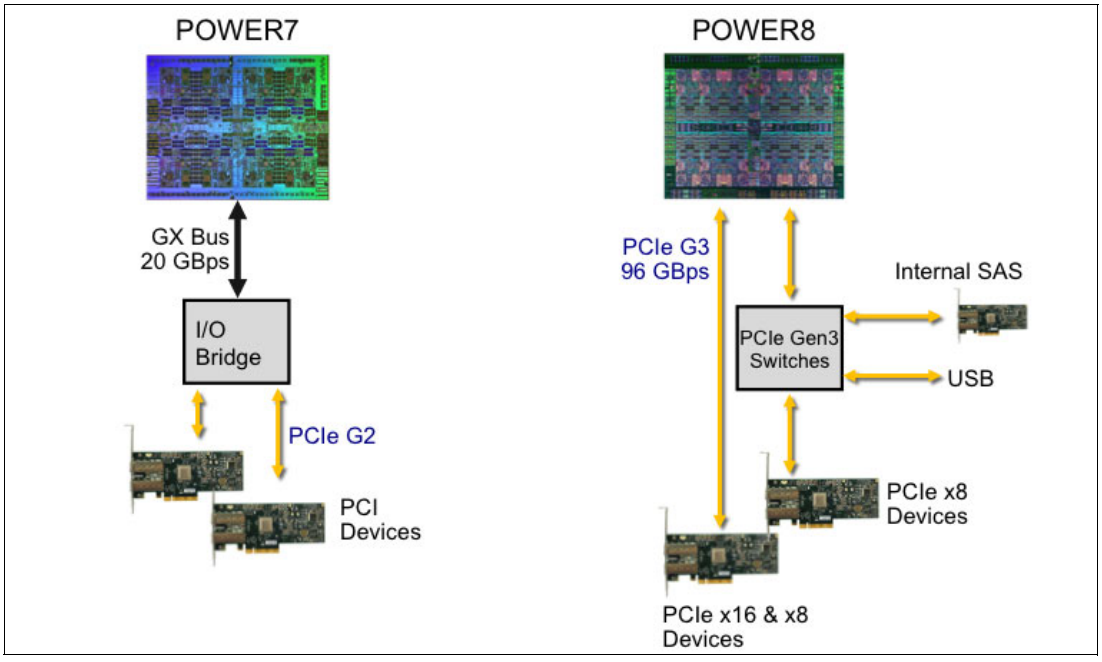


Figure 2-11 Comparison of POWER7 & POWER8 I/O buses architectures

Table 2-6 on page 40 lists the I/O bandwidth of Power S822 processor module configurations.

Table 2-6 I/O bandwidth

I/O	I/O bandwidth (maximum theoretical)
Total I/O bandwidth	Power S822 - one processor: 48 GBps simplex 96 GBps duplex Power S822 - two processors: 96 GBps simplex 192 GBps duplex

PCIe Interconnect: Each POWER8 processor has 48 PCIe lanes running at 8 Gbps full-duplex. The bandwidth formulas is calculates as follows:

```
48 lanes * 2 processors * 8 Gbps * 2 = 192 GBps
```

2.4 Internal I/O subsystem

The internal I/O subsystem resides on the system planar, which supports PCIe slots. PCIe adapters on the Power S822 are hot pluggable.

All PCIe slots support Enhanced Error Handling (EEH). PCI EEH-enabled adapters respond to a special data packet generated from the affected PCIe slot hardware by calling system firmware, which will examine the affected bus, allow the device driver to reset it, and continue without a system reboot. For Linux, EEH support extends to the majority of frequently used devices, although certain third-party PCI devices might not provide native EEH support.

2.4.1 Slot configuration

The amount of PCIe slots available on the Power S822 depends on the storage backplane used (#EJ0T or #EJ0U) and the amount of installed processors. The amount and speed of the available slots can be seen on the following table:

Table 2-7 PCIe slots versus server configuration

	One processor module	Two processor modules
Storage backplane #EJ0T	Qty 2 - x16 Gen3 low profile slots plus Qty 4 - x8 Gen3 low profile slots	Qty 4- x16 Gen3 low profile slots plus Qty 5- x8 Gen3 low profile slots
Storage backplane #EJ0U	Qty 2 - x16 Gen3 low profile slots plus Qty 3 - x8 Gen3 low profile slots	Qty 4- x16 Gen3 low profile slots plus Qty 4- x8 Gen3 low profile slots

Table 2-8 displays the PCIe Gen3 slot configuration for the server.

Table 2-8 Slot configuration of a Power S822

Slot	Description	Location code	Card size	Installed processors required to enable
Slot 1	PCIe Gen3 x8	P1-C2	Low profile	2
Slot 2	PCIe Gen3 x16	P1-C3	Low profile	2
Slot 3	PCIe Gen3 x16	P1-C5	Low profile	2
Slot 4	PCIe Gen3 x16	P1-C6	Low profile	1

Slot	Description	Location code	Card size	Installed processors required to enable
Slot 5	PCIe Gen3 x16	P1-C7	Low profile	1
Slot 6[a]	PCIe Gen3 x8	P1-C9	Low profile	1
Slot 7[b]	PCIe Gen3 x8	P1-C10	Low profile	1
Slot 8	PCIe Gen3 x8	P1-C11	Low profile	1
Slot 9	PCIe Gen3 x8	P1-C12	Low profile	1

a. Slot 6 (P1-C9) becomes obstructed by external SAS Ports and cannot be used by PCIe devices when #EJ0U is present. If a PCIe adapter is plugged into slot 6, then #EJ0U cannot be used or adapter must be moved to another suitable slot.

b. Included on all base configurations this slot (P1-C10) comes populated with a ethernet adapter PCIe2 LP 4-port 1 Gb Ethernet Adapter (#5260).

The Figure 2-12 shows the back view diagram of the server with the respective slot numbers.

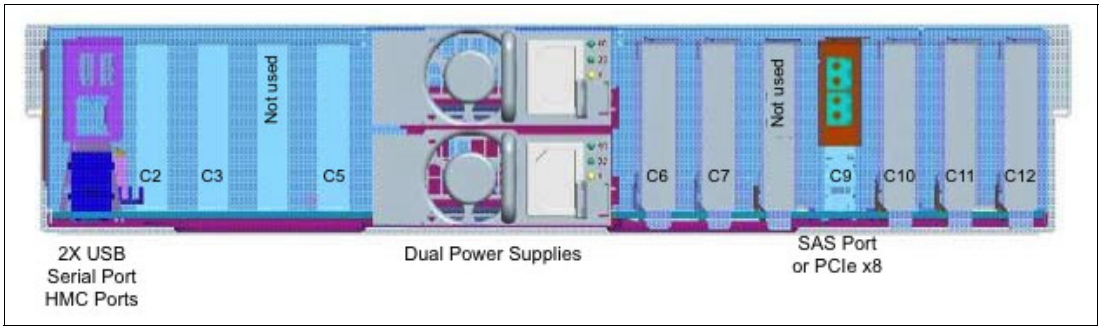


Figure 2-12 Back view diagram for Power S822

On Figure 2-13 on page 42 and Figure 2-14 on page 42 there are two top view diagrams showing the available PCIe slots on the two socket server considering both backplane options. The quantity of disk slots can vary depending on the storage backplane selected. For more details see2.6, “Internal storage” on page 51.

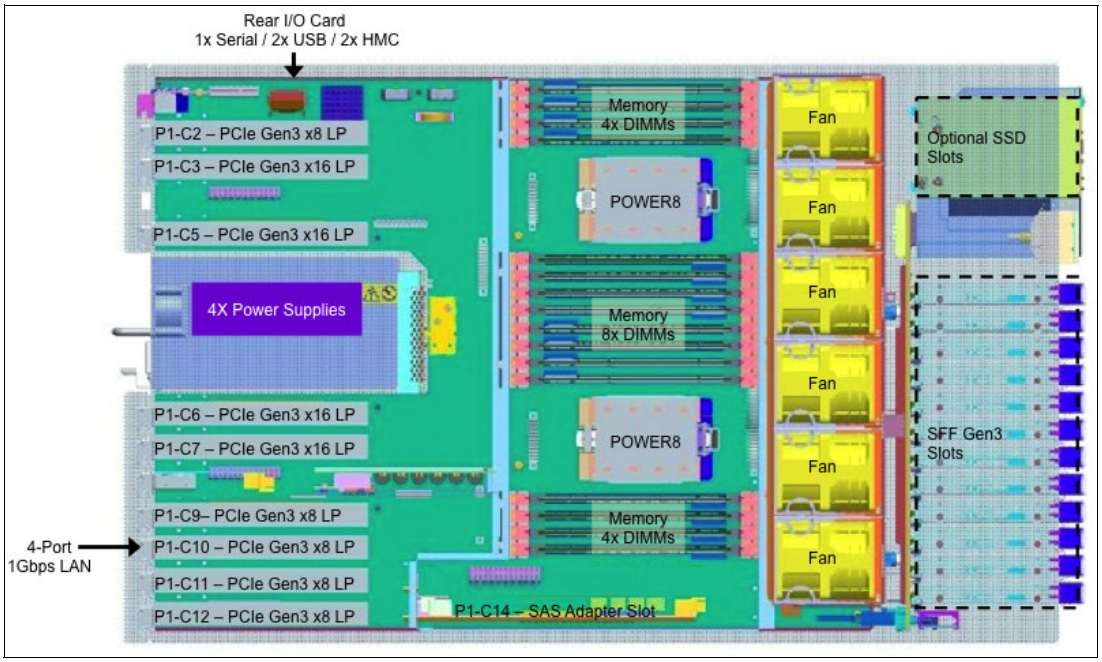


Figure 2-13 Top view diagram for a two socket Power S822 with #EJ0T backplane option

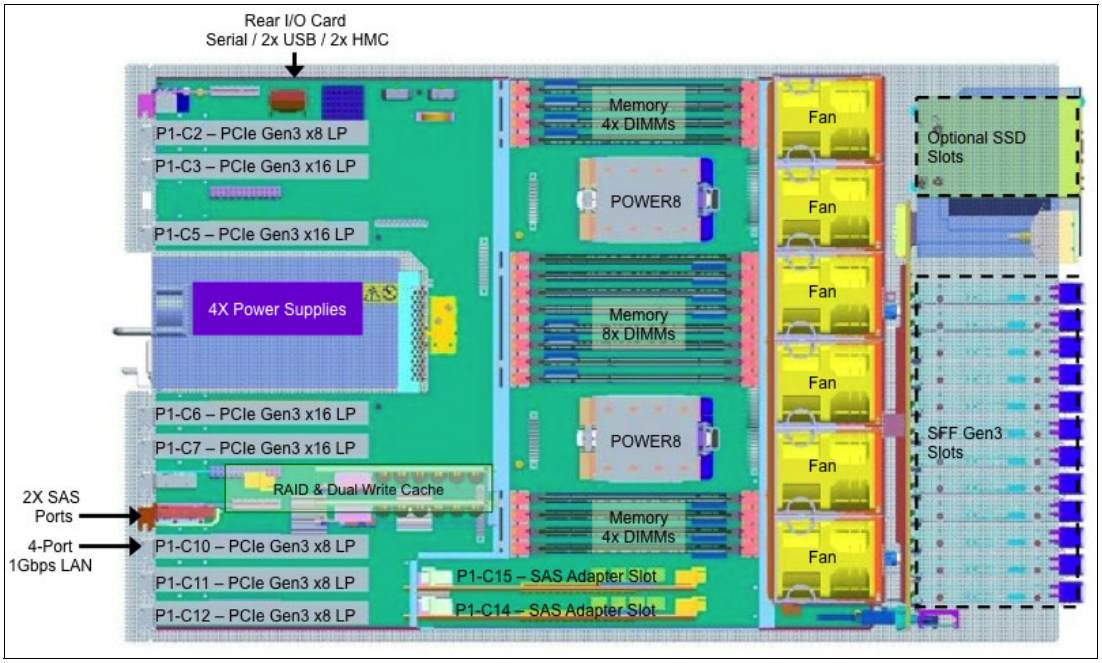


Figure 2-14 Top view diagram for a two socket Power S822 with #EJ0U backplane option

At least one PCIe Ethernet adapter is required on the server and therefore one of the x8 PCIe slots is used for this required adapter, identified as the P1-C10 slot. Included on all base configurations, as of the time of writing, this adapter is the PCIe2 LP 4-port 1 Gb Ethernet Adapter (#5260). It can be replaced or moved in field or as part of an upgrade, as long as the server still contains at least one ethernet adapter.

Remember: Slot 7 (P1-C10) comes with PCIe2 LP 4-port 1 Gb Ethernet Adapter (#5260) installed. In case this slot is needed, the adapter must be moved to another suitable slot or replaced by another ethernet adapter.

2.4.2 System ports

The system planar has one serial port that is called system port. When an HMC is connected to the server, the integrated system port of the server is rendered non-functional. In this case, you must install an asynchronous adapter, which is described in Table 2-16 on page 49, for serial port usage:

- Integrated system port is not supported under AIX or Linux when the HMC ports are connected to an HMC. Either the HMC ports, or the integrated system port can be used, but not both.
- The integrated system port is supported for modem and asynchronous terminal connections. Any other application using serial ports requires a serial port adapter to be installed in a PCI slot. The integrated system ports do not support IBM PowerHA configurations.
- Configuration of the integrated system port, including basic port settings (baud rate, and so on), modem selection, call-home and call-in policy, can be performed with the Advanced Systems Management Interface (ASMI).

Remember: The integrated console/modem port usage just described is for systems configured as a single, system-wide partition. When the system is configured with multiple partitions, the integrated console/modem ports are disabled because the TTY console and call home functions are performed with the HMC.

2.5 PCI adapters

This section covers the various types and functionality of the PCI adapters supported with the IBM Power S822 system.

2.5.1 PCI express

Peripheral Component Interconnect Express (PCIe) uses a serial interface and allows for point-to-point interconnections between devices (using a directly wired interface between these connection points). A single PCIe serial link is a dual-simplex connection that uses two pairs of wires, one pair for transmit and one pair for receive, and can transmit only one bit per cycle. These two pairs of wires are called a *lane*. A PCIe link can consist of multiple lanes. In such configurations, the connection is labeled as x1, x2, x8, x12, x16, or x32, where the number is effectively the number of lanes.

The PCIe interfaces supported on this server are PCIe Gen3, capable of 16 GBps simplex (32 GBps duplex) on a single x16 interface. PCIe Gen3 slots also support previous generations (Gen2 and Gen1) adapters, which operate at lower speeds, according to the following rules:

- Place x1, x4, x8, and x16 speed adapters in same connector size slots first, before mixing adapter speed with connector slot size.

- Adapters with smaller speeds are allowed in larger sized PCIe connectors but larger speed adapters are not compatible in smaller connector sizes (i.e. a x16 adapter cannot go in an x8 PCIe slot connector).

All adapters support Enhanced Error Handling (EEH). PCIe adapters use a different type of slot than PCI adapters. If you attempt to force an adapter into the wrong type of slot, you might damage the adapter or the slot.

IBM POWER8 processor based servers are able to support two different form factors of PCIe adapters:

- PCIe low profile (LP) cards, which are used with the Power S822 PCIe slots.
- PCIe full height and full high cards are not compatible with Power S822 and are designed to the following servers:
  - Power S814
  - Power S824

Before adding or rearranging adapters, use the System Planning Tool to validate the new adapter configuration. See the System Planning Tool website:

http://www.ibm.com/systems/support/tools/systemplanningtool/

If you are installing a new feature, ensure that you have the software required to support the new feature and determine whether there are any existing update prerequisites to install. To do this, use the IBM prerequisite website:

https://www-912.ibm.com/e_dir/eServerPreReq.nsf

The following sections discuss the supported adapters and provide tables of orderable feature numbers. The tables indicate operating system support (AIX, and Linux) for each of the adapters.

2.5.2 LAN adapters

In order to connect the Power S822 to a local area network (LAN), you can use the LAN adapters that are supported in the PCIe slots of the system unit. Table 2-9 lists the available LAN adapters.

Table 2-9 Available LAN adapters

Feature code	CCIN	Description	Maximum with one processor installed	Maximum with two processors installed	OS support
5260	576F	PCIe2 LP 4-port 1 GbE Adapter	6	8	AIX, Linux
5270	2B3B	PCIe LP 10 Gb FCoE 2-port Adapter	5	8	AIX, Linux
5271	5271	PCIe LP 4-Port 10/100/1000 Base-TX Ethernet Adapter	5	8	AIX, Linux
5272	5272	PCIe LP 10 GbE CX4 1-port Adapter	5	8	AIX, Linux
5274	5768	PCIe LP 2-Port 1 GbE SX Adapter	5	8	AIX, Linux

Feature code	CCIN	Description	Maximum with one processor installed	Maximum with two processors installed	OS support
5275	5275	PCIe LP 10 GbE SR 1-port Adapter	5	8	AIX, Linux
5279	2B53 2B43	PCIe2 LP 4-Port 10 GbE&1 GbE SFP+ Copper&RJ45	5	8	Linux
5280	2B54 2B44	PCIe2 LP 4-Port 10 GbE&1 GbE SR&RJ45 Adapter	5	8	Linux
5281	5767	PCIe LP 2-Port 1 GbE TX Adapter	5	8	AIX, Linux
5284	5287	PCIe2 LP 2-port 10 GbE SR Adapter	2	2	AIX, Linux
5286	5288	PCIe2 LP 2-Port 10 GbE SFP+ Copper Adapter	2	2	AIX, Linux
EC27	EC27	PCIe2 LP 2-Port 10 GbE RoCE SFP+ Adapter	5	8	AIX, Linux
EC29	EC29	PCIe2 LP 2-Port 10 GbE RoCE SR Adapter	0	8	AIX, Linux
EC2G		PCIe2 LP 2-port 10 GbE SFN6122F Adapter	5	8	Linux
EC2H		PCIe2 LP 2-port 10 GbE SFN5162F Adapter	5	8	Linux
EC3A	57BD	PCIe3 LP 2-Port 40 GbE NIC RoCE QSFP+ Adapter	5	8	AIX, Linux
EN0J	2B93	PCIe2 LP 4-port (10 Gb FCoE & 1 GbE) SR&RJ45	5	8	AIX, Linux
EN0L		PCIe2 LP 4-port(10 Gb FCoE & 1 GbE) SFP+Copper&RJ45	5	8	AIX, Linux
EN0N		PCIe2 LP 4-port(10 Gb FCoE & 1 GbE) LR&RJ45 Adapter	5	8	AIX, Linux
EN0T		PCIe2 LP 4-Port (10 Gb+1 GbE) SR+RJ45 Adapter	5	8	AIX, Linux
EN0V		PCIe2 LP 4-port (10 Gb+1 GbE) Copper SFP+RJ45 Adapter	5	8	AIX, Linux
EN0X		PCIe2 LP 2-port 10/1 GbE BaseT RJ45 Adapter	5	8	AIX, Linux

2.5.3 Graphics accelerator adapters

Table 2-10 on page 46 lists the available graphics accelerator adapters. The adapter can be configured to operate in either 8-bit or 24-bit color modes. The adapter supports both analog and digital monitors.

Table 2-10 Available graphics accelerator adapters

Feature code	CCIN	Description	Maximum with one processor installed	Maximum with two processors installed	OS support
5269	5269	PCIe LP POWER GXT145 Graphics Accelerator	5	6	AIX, Linux

2.5.4 SAS adapters

2.5.5, “PCIe RAID and SSD SAS adapter” on page 46 lists the SAS adapter that are available for Power S822 systems

Table 2-11 Available SAS adapters

Feature code	CCIN	Description	Maximum with one processor installed	Maximum with two processors installed	OS support
5278	2D1F	PCIe LP 2-x4-port SAS Adapter 3Gb	2	2	AIX, Linux
ESA2	57C4	PCIe2 LP RAID SAS Adapter Dual-port 6 Gb	5	8	AIX, Linux

2.5.5 PCIe RAID and SSD SAS adapter

Supported on selected POWER8 processor-based servers the PCIe RAID and SSD SAS adapter has up to four 177 GB SSD (#1995) modules, plugged directly onto the adapter, providing additional SSD capacity, saving the need for the SAS bays and cabling.

Figure 2-15 shows the double-wide adapter and SSD modules.

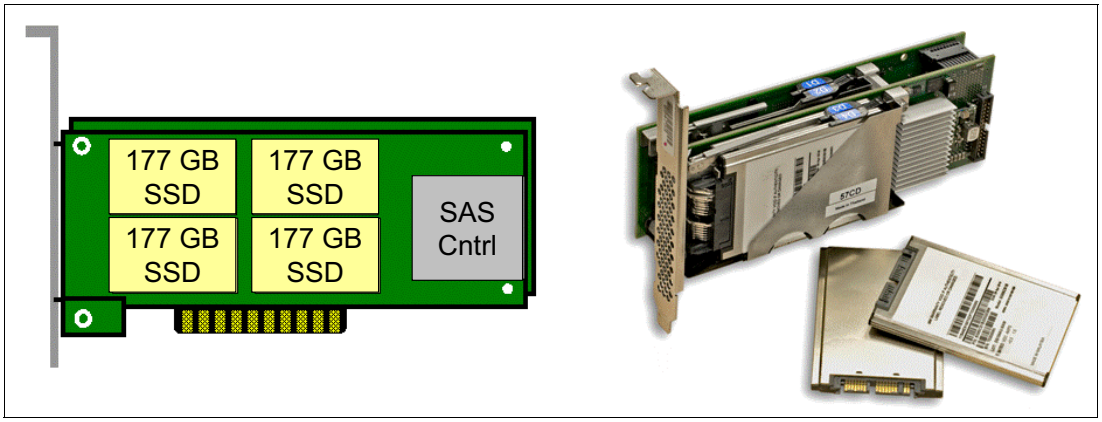


Figure 2-15 The PCIe RAID and SSD SAS adapter, and 177 GB SSD modules

Table 2-12 shows available RAID and SSD SAS adapters for the Power S822.

Table 2-12 Available SSD SAS adapters.

Feature code	CCIN	Description	Maximum with one processor installed	Maximum with two processors installed	OS support
2053	57CD	PCIe LP RAID & SSD SAS Adapter 3Gb	1	1	AIX, Linux

The 177 GB SSD module with enterprise multi-level cell (eMLC) uses an enterprise-class MLC flash technology, which provides enhanced durability, capacity, and performance. One, two, or four modules can be plugged onto a PCIe RAID and SSD SAS adapter, providing up to 708 GB of SSD capacity on one PCIe adapter.

Because the SSD modules are mounted on the adapter, to service either the adapter or one of the modules, the entire adapter must be removed from the system.

Under AIX and Linux, the 177 GB modules can be reformatted as JBOD disks, providing 200 GB of available disk space. This way removes RAID error correcting information, so the best approach to prevent data loss in case of failure is to mirror the data by using operating system tools.

2.5.6 Fibre Channel adapters

The systems support direct or SAN connection to devices that use Fibre Channel adapters. Table 2-13 summarizes the available Fibre Channel adapters, which all have LC connectors.

If you are attaching a device or switch with an SC type fiber connector, then an LC-SC 50 Micron Fiber Converter Cable (#2456) or an LC-SC 62.5 Micron Fiber Converter Cable (#2459) is required.

Table 2-13 Available Fibre Channel adapters

Feature code	CCIN	Description	Maximum with one processor installed	Maximum with two processors installed	OS support
5273	577D	PCIe LP 8 Gb 2-Port Fibre Channel Adapter	5	8	AIX, Linux
5276	5774	PCIe LP 4 Gb 2-Port Fibre Channel Adapter	5	8	AIX, Linux
EN0B	577F	PCIe2 LP 16 Gb 2-port Fibre Channel Adapter	5	8	AIX, Linux
EN0Y	EN0Y	PCIe2 LP 8 Gb 4-port Fibre Channel Adapter	5	8	AIX, Linux

NPIV: The use of N_Port ID Virtualization (NPIV) through the Virtual I/O server requires an NPIV-capable Fibre Channel adapter, such as the #5273, #EN0B and #EN0Y.

2.5.7 Fibre Channel over Ethernet

Fibre Channel over Ethernet (FCoE) allows for the convergence of Fibre Channel and Ethernet traffic onto a single adapter and a converged fabric.

Figure 2-16 compares existing Fibre Channel and network connections and FCoE connections.

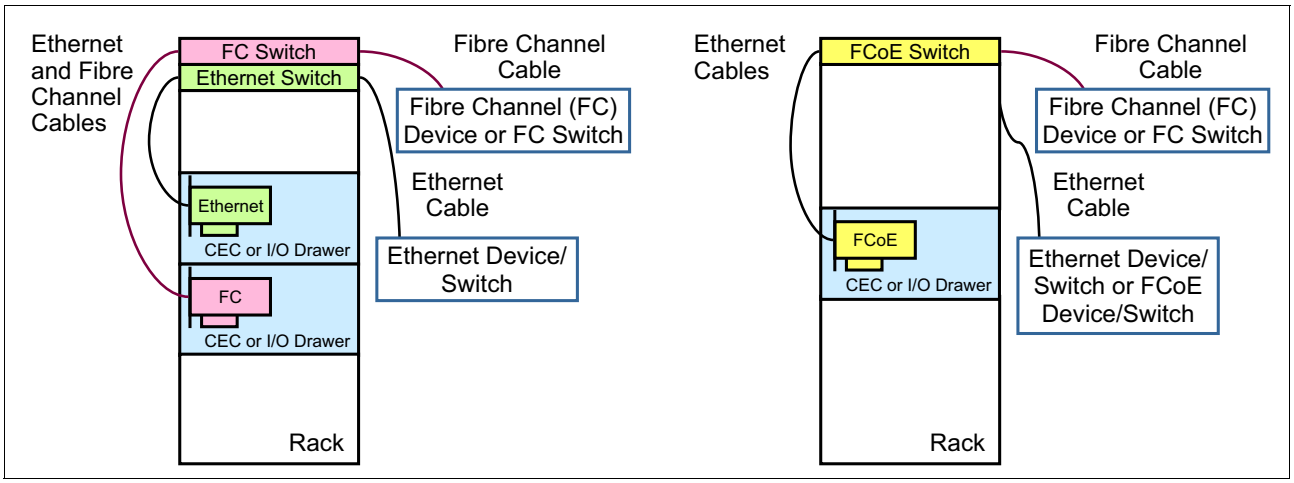


Figure 2-16 Comparison between existing Fibre Channel and network connection and FCoE connection

Table 2-13 on page 47 lists the available Fibre Channel over Ethernet Adapter. It is a high-performance, Converged Network Adapter (CNA) using SR optics. Each port can simultaneously provide network interface card (NIC) traffic and Fibre Channel functions.

Table 2-14 Available FCoE adapters

Feature code	CCIN	Description	Maximum with one processor installed	Maximum with two processors installed	OS support
5270	2B3B	PCIe LP 10 Gb FCoE 2-port Adapter	5	8	AIX, Linux
EN0J	2B93	PCIe2 LP 4-port (10 Gb FCoE & 1 GbE) SR&RJ45	5	8	AIX, Linux
EN0L		PCIe2 LP 4-port (10 Gb FCoE & 1 GbE) SFP+Copper&RJ45	5	8	AIX, Linux
EN0N		PCIe2 LP 4-port (10 Gb FCoE & 1 GbE) LR&RJ45 Adapter	5	8	AIX, Linux

For more information about FCoE, read *An Introduction to Fibre Channel over Ethernet, and Fibre Channel over Convergence Enhanced Ethernet*, REDP-4493.

2.5.8 InfiniBand Host Channel adapter

The InfiniBand Architecture (IBA) is an industry-standard architecture for server I/O and inter-server communication. It was developed by the InfiniBand Trade Association (IBTA) to provide the levels of reliability, availability, performance, and scalability necessary for present and future server systems with levels significantly better than can be achieved using bus-oriented I/O structures.

InfiniBand (IB) is an open set of interconnect standards and specifications. The main IB specification is published by the InfiniBand Trade Association and is available at:

http://www.infinibandta.org/

InfiniBand is based on a switched fabric architecture of serial point-to-point links, where these IB links can be connected to either host channel adapters (HCAs), used primarily in servers, or target channel adapters (TCAs), used primarily in storage subsystems.

The InfiniBand physical connection consists of multiple byte lanes. Each individual byte lane is a four-wire, 2.5, 5.0, or 10.0 Gbps bidirectional connection. Combinations of link width and byte lane speed allow for overall link speeds from 2.5 Gbps to 120 Gbps. The architecture defines a layered hardware protocol and also a software layer to manage initialization and the communication between devices. Each link can support multiple transport services for reliability and multiple prioritized virtual communication channels.

For more information about InfiniBand, see *HPC Clusters Using InfiniBand on IBM Power Systems Servers*, SG24-7767.

A connection to supported InfiniBand switches is accomplished by using the QDR optical cables #3290 and #3293.

Table lists the available InfiniBand adapters.

Table 2-15 Available Infiniband adapters

Feature code	CCIN	Description	Maximum with one processor installed	Maximum with two processors installed	OS support
5283	58E2	PCIe2 LP Dual-Port 4X IB QDR adapter 40 Gb	5	8	AIX, Linux

2.5.9 Asynchronous and USB adapters

Asynchronous PCIe adapters provide connection of asynchronous EIA-232 or RS-422 devices. If you have a cluster configuration or high-availability configuration and plan to connect the IBM Power Systems using a serial connection, you can use the features listed in Table 2-16.

Table 2-16 Available Asynchronous and USB adapters

Feature code	CCIN	Description	Maximum with one processor installed	Maximum with two processors installed	OS support
EC45		PCIe2 LP 4-Port USB 3.0 Adapter	5	8	AIX, Linux

Feature code	CCIN	Description	Maximum with one processor installed	Maximum with two processors installed	OS support
5227	57D2	PCIe LP 4-Port Async EIA-232 adapter	5	8	AIX, Linux
5290	57D4	PCIe LP 2-Port Async EIA-232 adapter	5	8	AIX, Linux

2.5.10 Cryptographic coprocessor

The cryptographic coprocessor card is not available to the Power S822 due to the adapter size.

2.5.11 FPGA Adapters

The field-programmable gate array (FPGA) adapters are PCIe adapters based on a semiconductor device that can be programmed.

Unlike an ASIC, which is designed and programmed to perform a single function, an FPGA can be programmed to execute different product functions, adapt to new standards, and reconfigure its hardware for specific applications even after the product has been installed in the field. An FPGA is able to implement any logical function that an application-specific integrated circuit (ASIC) would perform, but has the ability to have its code update to include more functions or perform a different role.

Today FPGAs have not only the logic elements, capable of being reprogrammed, but also include SRAM memory, high-speed interconnects that can range up to 400 Gbps, logic blocks and routing. This allows for very versatile solutions once a single adapter is able to perform several distinct functions depending on the code deployed on it.

By having a highly optimized software stack, the FPGAs can act as coprocessor for the server CPUs, executing repetitive and complex functions at a fraction of the time and power, while allowing for the server CPUs to perform other functions at the same time.

The FPGA adapter cards supported for the Power S822 can be seen on Table 2-17.

Table 2-17 Available FPGA adapters

Feature code	CCIN	Description	One Processor	Two Processors	OS support
EJ13		PCIe3 LP FPGA Accelerator Adapter	1	1	AIX

The #EJ13 is a low profile FPGA adapter based on Altera Stratix V 28 nm hardware. Besides all the logic components it has also 8 GB DDR3 RAM.

The initial implementation for the #EJ13 is an adapter capable of performing gzip compression. The zlib API, a software library responsible for data compression used in several software, is capable of moving the tasks directly to the FPGA allowing for increased compression performance, increased compression rates and decreased CPU usage. JAVA 7.1 is already enabled to that take advantage of this kind of acceleration.

Other applications like Big Data can take advantage of this approach on compression once the compressions rates are higher than the ones achieved via software and the compression processes complete in a shorter time allowing for more data density, disk savings and faster data analysis.

2.6 Internal storage

The internal storage on the Power S822 server depends on the DASD/Media backplane used. The server supports two DASD/Media backplanes: #EJ0T and #EJ0U.

The #EJ0T storage backplane contains:

- A storage backplane for twelve 2.5-inch SFF Gen3 HDDs or SSDs.
- One SAS disk controller capable of RAID 0, RAID 5, RAID 6, and RAID 10, placed in a dedicated SAS controller slot (P1-C14).
- The optional split backplane feature #EJ0V adds a secondary SAS disk controller and allows DASDs to be split into two groups of six (6+6). This secondary controller is placed in the second dedicated SAS controller slot (P1-C15).

The #EJ0U storage backplane contains:

- A storage backplane for eight 2.5-inch SFF Gen3 HDDs or SSDs.
- Two active-active SAS disk controllers capable of RAID 0, RAID 5, RAID 6, RAID 10, RAID 5T2, RAID 6T2 & RAID 10T2, placed in dedicated SAS controller slots P1-C14 and P1-C15.
- The #EJTL SSD Module Cage, automatically added by e-config when #EJ0U is ordered, adds a secondary disk cage with six 1.8” SSD module bays.
- Two external SAS ports for DASD drawers connectivity (via slot P1-C9) supporting one EXP24S SFF Gen2-bay Drawer (#5887).
- The storage split backplane function is not supported.

Table 2-18 presents a summarized view of the above features.

Table 2-18 Backplane options and summary of features

	#EJ0T backplane	#EJ0U backplane
Supported RAID types	JBOD, RAID 0, RAID 5, RAID 6, RAID 10	JBOD, RAID 0, RAID 5, RAID 6, RAID 10, RAID 5T2, RAID 6T2
Disk bays	12 SFF Gen3 (HDDs/SSDs)	8 SFF Gen3 (HDDs/SSDs)
SAS controllers	Single	Dual active-active
Easy Tier capable controllers	No	Yes
External SAS ports	No	Yes, two SAS ports
Split backplane	Optional (#EJ0V) includes secondary SAS controller	No
SSD 1.8” module cage	No	Optional (#EJTL)
PCIe slot P1-C9	Available for PCIe x8 cards	Used by external SAS ports

The 2.5-inch or small form factor (SFF) SAS bays can contain SAS drives (HDD or SSD) mounted on a Gen3 tray or carrier (also knows as SFF-3). SFF-1 or SFF-2 drives do not fit in an SFF-3 bay. All SFF-3 bays support concurrent maintenance or “hot plug” capability.

Additionally, as an optional for the #EJ0U backplane, the feature #EJTL adds a 6-bay 1.8-inch SSD Cage behind the server bezel. All six bays are accessed by both of the SAS controllers and the bays support concurrent maintenance (hot plug). At the time of writing, the supported disk on these bays is feature #ES16, 1.8-inch SSD with 387 GB capacity.

The SSF-3 disk drives connect to the DASD backplane and are hot-swap and front-accessible while optional the 1.8-inch SSD driver are housed behind the server bezel but also hot-swap once you remove the front bezel.

Server front view considering both backplane option can be seen on the diagram on Figure 2-17.

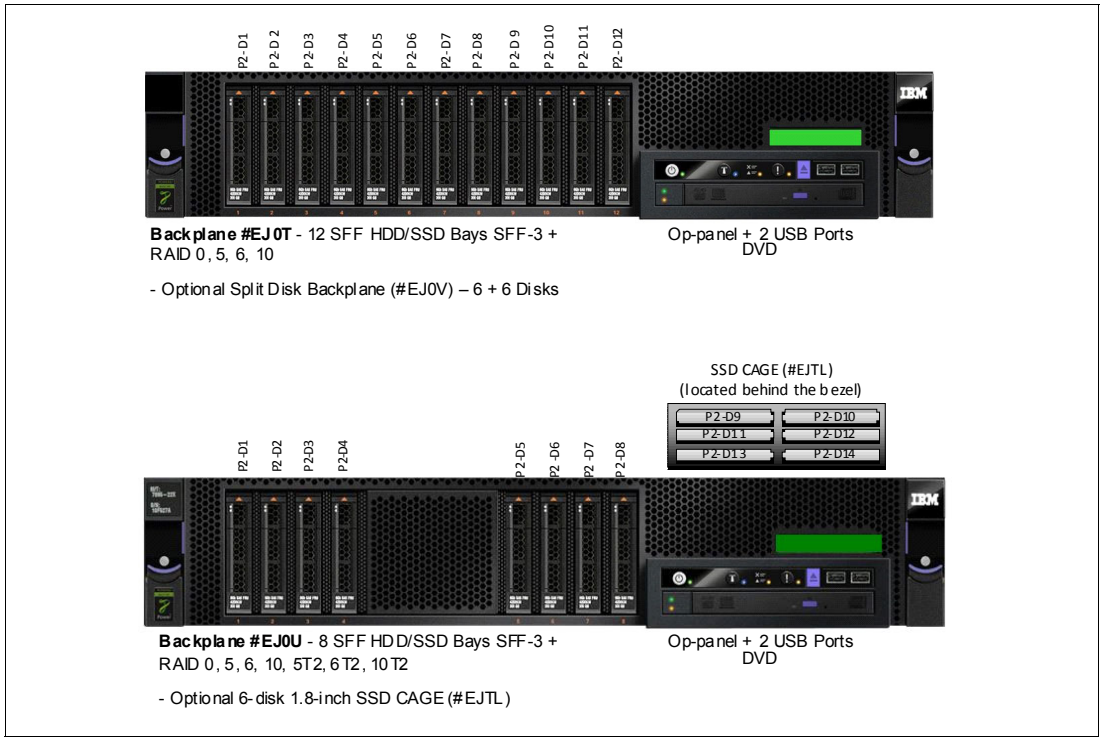


Figure 2-17 Server front view with different backplane options

The internal connections to the physical disks can be seen on the following diagrams.

Figure 2-18 shows the internal connections when using #EJ0T.

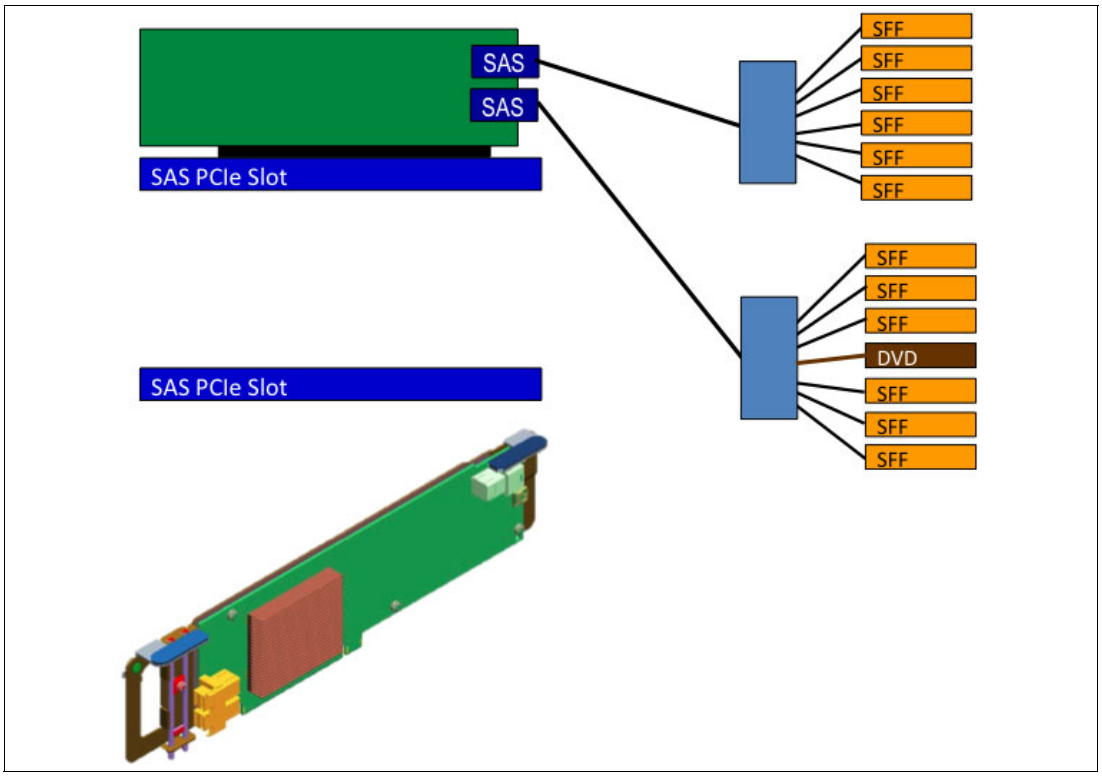


Figure 2-18 Internal topology overview for #EJ0T DASD backplane

Figure 2-19 shows the internal topology overview for the #EJ0T backplane in a split backplane configuration (optional #EJ0V).

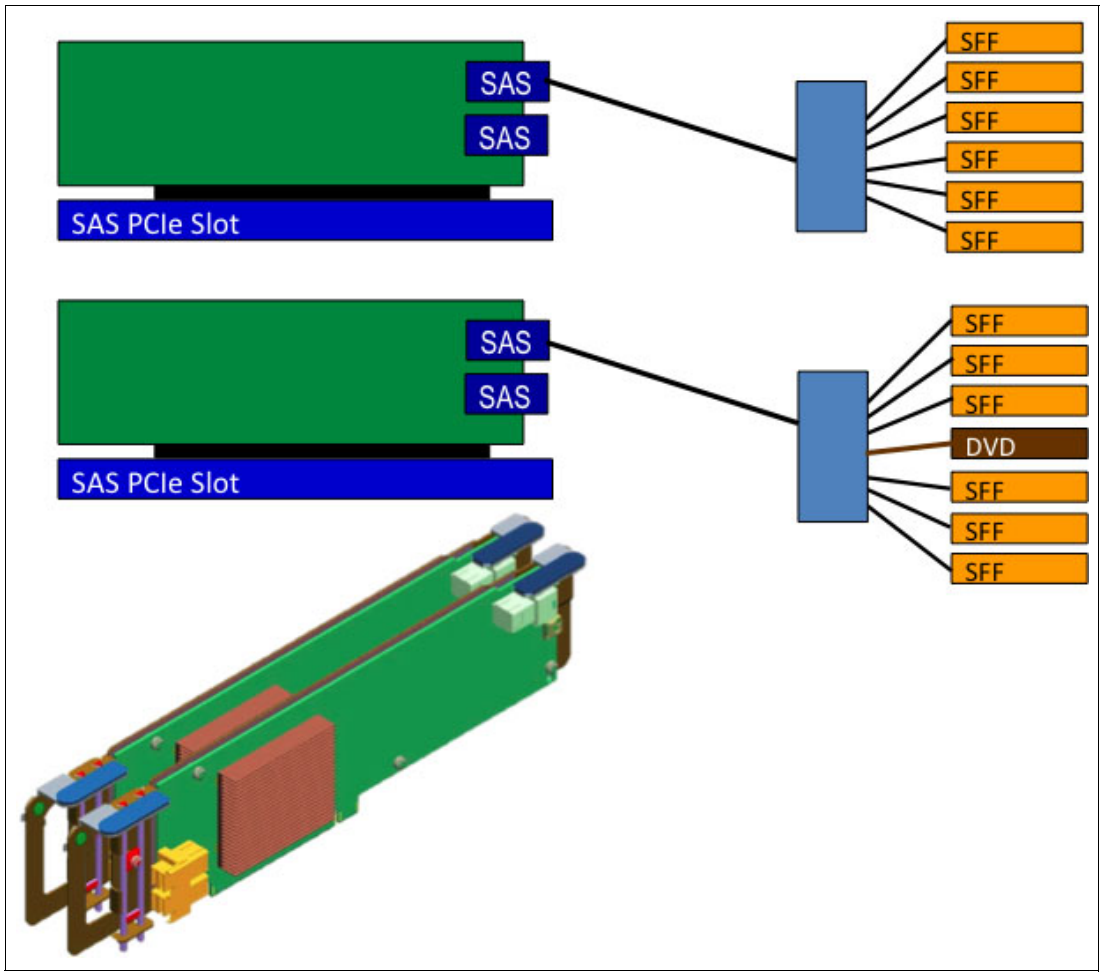


Figure 2-19 Internal topology overview for the #EJ0T DASD backplane with #EJ0V split backplane feature

Figure 2-20 shows the details of the internal topology overview for the #EJ0U DASD backplane with optional #EJTL 1.8-inch SSD Cage.

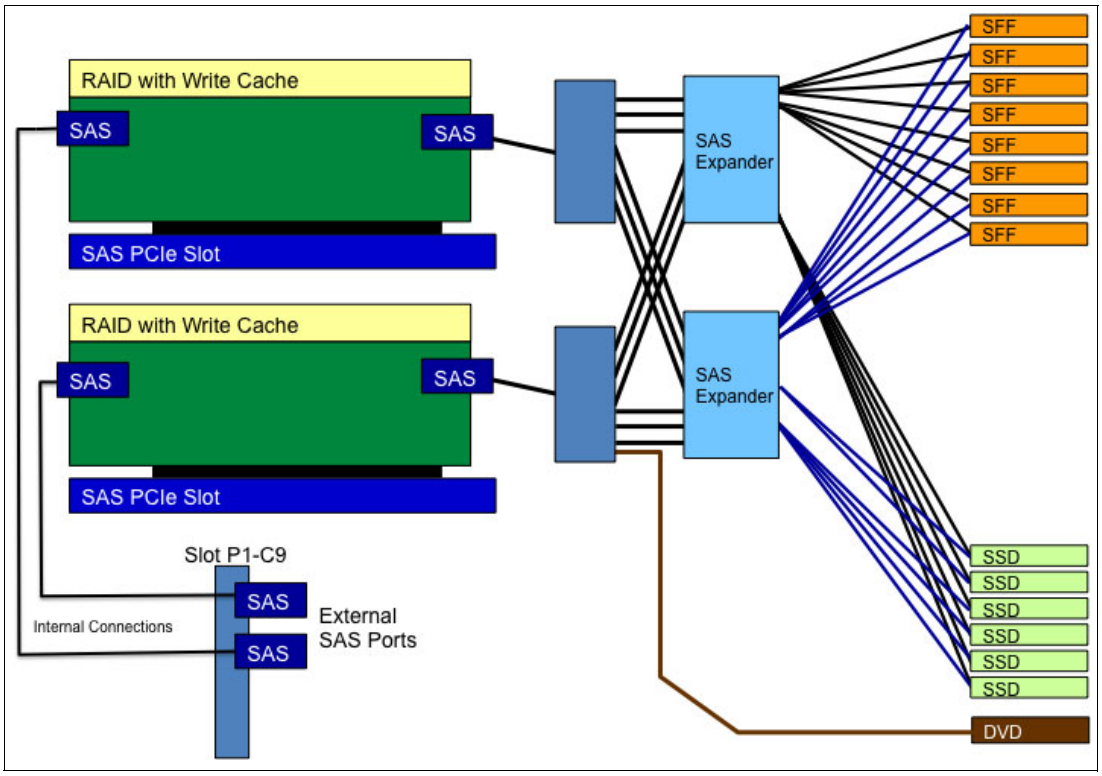


Figure 2-20 Internal topology overview for the #EJ0U DASD backplane

The external SAS ports provided by the #EJ0U support the attachment of a single EXP24S SFF Gen2-bay Drawer (#5887). Additional disk drawers can be attached by using supported PCIe SAS adapters.

2.6.1 RAID support

There are multiple protection options for HDD/SSD drives in the Power S822 systems, whether they are contained in the SAS SFF bays in the system unit or drives in disk-only I/O drawers. Although protecting drives is always recommended, AIX/Linux users can choose to leave a few or all drives unprotected at their own risk, and IBM supports these configurations.

Drive protection

HDD/SSD drive protection can be provided by the AIX, and Linux operating system, or by the HDD/SSD hardware controllers.

The default storage backplane #EJ0T contains one SAS HDD/SSD controller and provides support for JBOD and RAID 0, 5, 6, and 10 for AIX or Linux. A secondary non-redundant controller is added when using split backplane feature #EJ0V so each six disks have a separated disk controller.

This controller is replaced when choosing the optional #EJ0U storage backplane by a pair of High Performance RAID controllers with dual integrated SAS controllers with 1.8 GB of physical write cache. High performance controllers run SFF-3 SAS bays, 1.8-inch SSD cage bays, and DVD bay in the system unit. Dual controllers (also called dual I/O adapters or paired controllers) and their write cache are placed in integrated slots and do not use PCIe

slots. However cache power protection hardware covers one PCIe x8 slot (P1-C9). Patented Active/Active configurations with at least two arrays is supported.

The write cache, responsible for increasing write performance by caching data before it is written do the physical disks, can have its data compression capabilities activated, providing up to 7.2 GB effective cache capacity. The write cache contents are protected against power loss with flash memory and super capacitors removing the need for battery maintenance.

The high performance SAS controllers provide RAID 0, RAID 5, RAID 6 and RAID 10 support and its EasyTier variations (RAID 5T2, RAID 6T2, RAID 10T2) provided the server has both HDDs and SSDs installed.

Easy Tier function is supported so the dual controllers can automatically move hot data to attached SSD and cold data to attached HDD for AIX/Linux/VIOS environments. If a EXP24S SFF Gen2-bay Drawer (#5887) is attached to the adapters, Easy Tier funcionality is also extended to the disks on this drawer. To learn more about Easy Tier funcionality read 2.6.2, “Easy Tier” on page 57.

Table 2-19 lists the RAID support configurations by the storage backplane options.

Table 2-19 RAID support configurations

Storage backplane	JBOD	RAID 0, 5, 6, and 10	RAID 0, 5, 6 and 10 and Easy Tier (RAID 5T2, 6T2, 10T2)	Split backplane	External SAS port
#EJ0T	Yes	Yes	No	Optional	No
#EJ0U	Yes	Yes	Yes	No	Yes

AIX and Linux can use disk drives that are formatted with 512-byte blocks when being mirrored by the operating system. These disk drives must be reformatted to 528-byte sectors when used in RAID arrays. Although a small percentage of the drive's capacity is lost, additional data protection such as ECC and bad block detection is gained in this reformatting. For example, a 300 GB disk drive, when reformatted, provides approximately 283 GB. Solid-state drives are always formatted with 528 byte sectors.

Supported RAID functions

Base hardware supports RAID 0, 5, 6, and 10. When additional features are configured, the server supports hardware RAID 0, 5, 6, 10, 5T2, 6T2 and 10T2:

- RAID 0 provides striping for performance, but does not offer any fault tolerance.

  The failure of a single drive results in the loss of all data on the array. This version of RAID increases I/O bandwidth by simultaneously accessing multiple data paths.

- RAID 5 uses block-level data striping with distributed parity.

  RAID RAID 5 stripes both data and parity information across three or more drives. Fault tolerance is maintained by ensuring that the parity information for any given block of data is placed on a drive separate from those used to store the data itself. This version of RAID provides data resiliency in the case of a single drive failing in a RAID 5 array.

- RAID 6 uses block-level data striping with dual distributed parity.

  RAID 6 is the same as RAID 5 except that it uses a second level of independently calculated and distributed parity information for additional fault tolerance. RAID 6 configuration requires N+2 drives to accommodate the additional parity data, making it less cost effective than RAID 5 for equivalent storage capacity. This version of RAID provides data resiliency in the case of one or two drives failing in a RAID 6 array. It is

designed when working with large capacity disks, once it allows to sustain data parity during the rebuild process.

- RAID 10 is also known as a striped set of mirrored arrays.

  It is a combination of RAID 0 and RAID 1. A RAID 0 stripe set of the data is created across a two-disk array for performance benefits. A duplicate of the first stripe set is then mirrored on another two-disk array for fault tolerance. This version of RAID provides data resiliency in the case of a single drive failure and may provide resiliency for multiple drive failures.

- RAID 5T2, RAID 6T2 and RAID 10T2 are the same RAIDs as defined above but with EasyTier enabled. It requires that both types of disks exists on the system under the same controller (HDDs and SSDs) and that both are configured under the same RAID type.

2.6.2 Easy Tier

With the standard SAS adapter (#EJ0T) the server can handle both HDDs and SSDs attached to its storage backplane as long as they are on separate arrays.

The High Performance RAID adapters (#EJ0U) has the ability to handle both types of storage in two different ways:

- Separate Arrays: SSDs and HDDs co-exist on separate arrays, just like the Standard SAS Adapter would do.
- Easy Tier: SSDs and HDDs co-exist under the same array.

When under the same array, the adapter can automatically move most accessed data to faster storage (SSDs) and less accessed data to slower storage (HDDs). This is called Easy Tier.

There is no need of coding or software intervention once the RAID is properly configured. Statistics on block accesses are gathered every minute and once the adapter realizes that some portion of the data is being frequently requested, it will move this data to faster devices. The data is moved in chunks of 1 MB or 2 MB called “Bands”.

From the operating system point-of-view, there is just a regular Array Disk. From the SAS controller point-of-view, there are two arrays with parts of the data being serviced by one tier of disks and parts by another tier of disks.

Figure 2-21 on page 58 shows a diagram of an Easy Tier array.

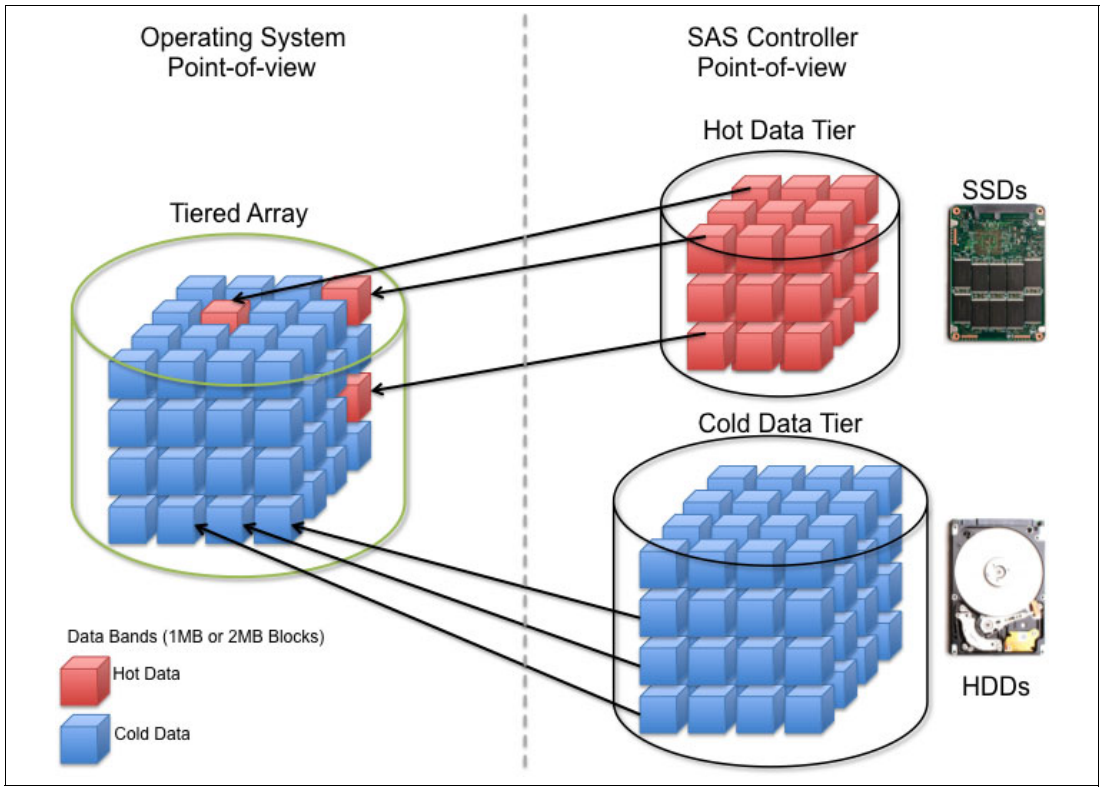


Figure 2-21 Easy Tier Diagram

The Easy Tier configuration is accomplished via standard operating system SAS adapter configuration utility. On Figure 2-22 and Figure 2-23 on page 59 there are two examples of tiered array creation for AIX.

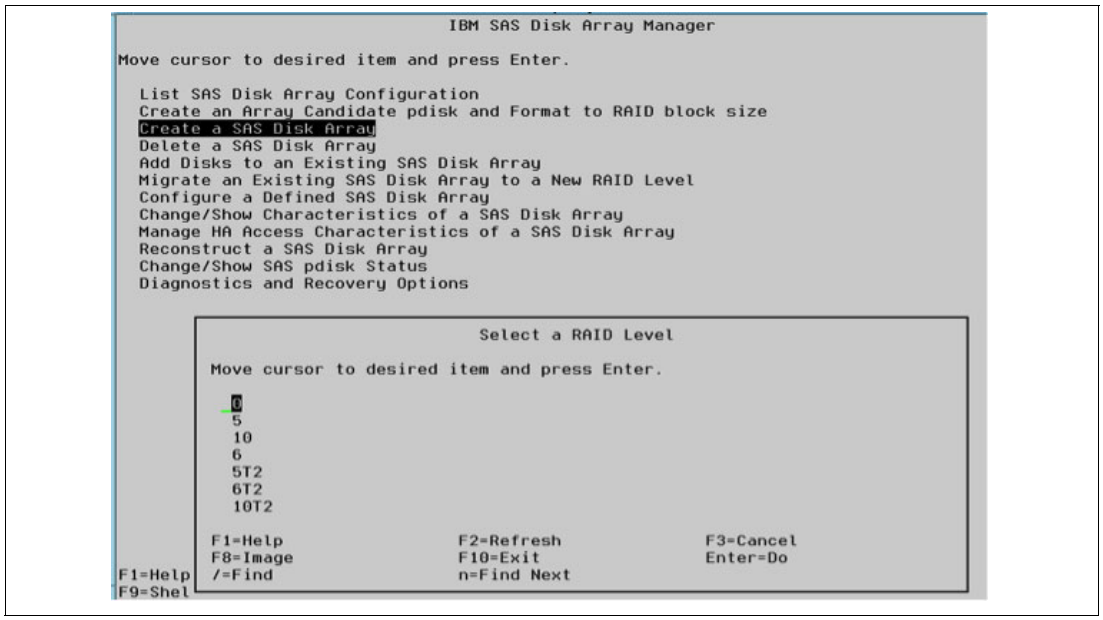

IBM SAS Disk Array Manager
Move cursor to desired item and press Enter.
List SAS Disk Array Configuration
Create an Array Candidate pdisk and Format to RAID block size
Create a SAS Disk Array
Delete a SAS Disk Array
Add Disks to an Existing SAS Disk Array
Migrate an Existing SAS Disk Array to a New RAID Level
Configure a Defined SAS Disk Array
Change/Show Characteristics of a SAS Disk Array
Manage HA Access Characteristics of a SAS Disk Array
Reconstruct a SAS Disk Array
Change/Show SAS pdisk Status
Diagnostics and Recovery Options
Select a RAID Level
Move cursor to desired item and press Enter.
0
5
10
6
5T2
6T2
10T2
F1=Help F2=Refresh F3=Cancel
F8=Image F10=Exit Enter=Do
/=Find n=Find Next
F1=Help
F9=Shel

Figure 2-22 Array type selection screen on AIX RAID Manager

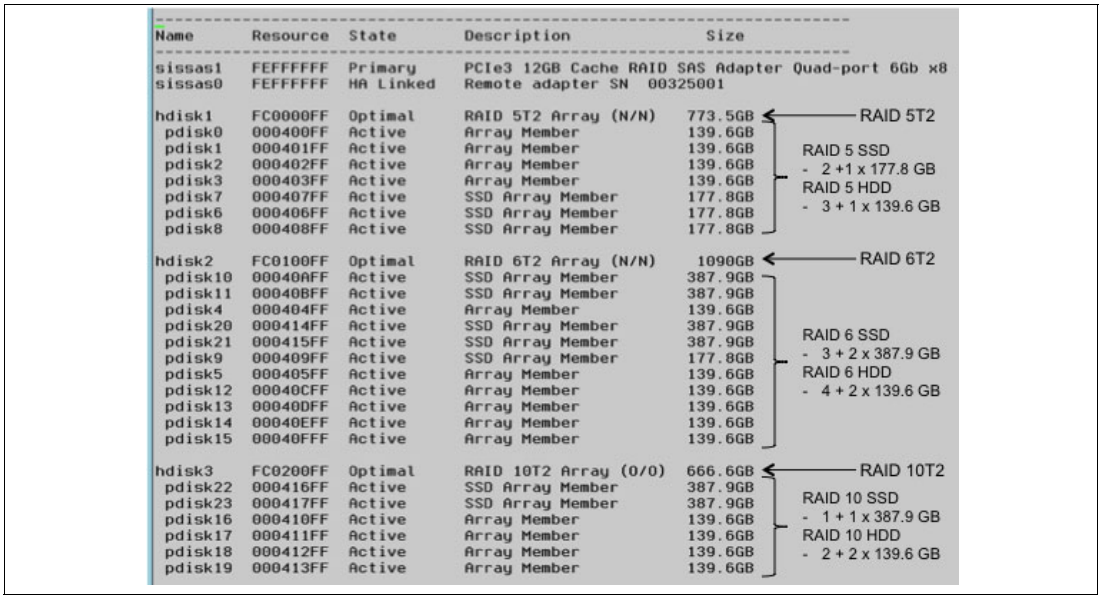

Name Resource State Description Size
sissas1 FEFFFFFF Primary PCIe3 12GB Cache RAID SAS Adapter Quad-port 6Gb x8
sissas0 FEFFFFFF HA Linked Remote adapter SN 00325001
hdisk1 FC0000FF Optimal RAID 5T2 Array (N/N) 773.5GB
pdisk0 000400FF Active Array Member 139.6GB
pdisk1 000401FF Active Array Member 139.6GB
pdisk2 000402FF Active Array Member 139.6GB
pdisk3 000403FF Active Array Member 139.6GB
pdisk7 000407FF Active SSD Array Member 177.8GB
pdisk6 000406FF Active SSD Array Member 177.8GB
pdisk8 000408FF Active SSD Array Member 177.8GB
hdisk2 FC0100FF Optimal RAID 6T2 Array (N/N) 1090GB
pdisk10 00040AFF Active SSD Array Member 387.9GB
pdisk11 00040BFF Active SSD Array Member 387.9GB
pdisk4 000404FF Active Array Member 139.6GB
pdisk20 000414FF Active SSD Array Member 387.9GB
pdisk21 000415FF Active SSD Array Member 387.9GB
pdisk9 000409FF Active SSD Array Member 177.8GB
pdisk5 000405FF Active Array Member 139.6GB
pdisk12 00040CFF Active Array Member 139.6GB
pdisk13 00040DFF Active Array Member 139.6GB
pdisk14 00040EFF Active Array Member 139.6GB
pdisk15 00040FFF Active Array Member 139.6GB
hdisk3 FC0200FF Optimal RAID 10T2 Array (0/0) 666.6GB
pdisk22 000416FF Active SSD Array Member 387.9GB
pdisk23 000417FF Active SSD Array Member 387.9GB
pdisk16 000410FF Active Array Member 139.6GB
pdisk17 000411FF Active Array Member 139.6GB
pdisk18 000412FF Active Array Member 139.6GB
pdisk19 000413FF Active Array Member 139.6GB
RAID 5T2
RAID 5 SSD
- 2 +1 x 177.8 GB
RAID 5 HDD
- 3 + 1 x 139.6 GB
RAID 6T2
RAID 6 SSD
- 3 + 2 x 387.9 GB
RAID 6 HDD
- 4 + 2 x 139.6 GB
RAID 10T2
RAID 10 SSD
- 1 + 1 x 387.9 GB
RAID 10 HDD
- 2 + 2 x 139.6 GB

Figure 2-23 Tiered Arrays (RAID 5T2, RAID 6T2 and RAID 10T2) example on AIX Raid Manager

In order to support Easy Tier make sure the server is running at least the minimum versions described below:

- VIOS 2.2.3.3 with interim fix IV56366 or later
- AIX V7.1 TL3 SP3 or later
- AIX V6.1 TL9 SP3 or later
- RHEL 6.5 or later
- SLES 11 SP3 or later

2.6.3 External SAS port

The Power S822 DASD backplane (#EJ0U) offers a connection to an external SAS port:

- The SAS port connector is located on slot P1-C9.
- The external SAS port is used for expansion to one external SAS EXP24S SFF Gen2-bay Drawer (#5887).

Additional drawers and the IBM System Storage 7226 Tape and DVD Enclosure Express (Model 1U3) can be attached by installing additional SAS adapters.

Note: Only one SAS drawer is supported from the external SAS port. Additional SAS drawers can be supported through SAS adapters. SSDs are not supported on the SAS drawer connected to the external port.

2.6.4 Media bays

Included in the feature EJ0T or EJ0U backplanes there is a slimline media bay that can optionally house a SATA DVD-RAM (#5771). Direct dock and hot-plug of the DVD media device is supported.

The DVD drive and media device do not have an independent SAS adapter and so cannot be assigned to an LPAR independently of the HDD/SSDs in the system.

2.7 External I/O subsystems

At the time of writing there are no PCIe I/O drawers such as the feature 5802, or 5877 supported on the Power S822.

Statement of direction: IBM plans to introduce an PCIe I/O drawer which will expand the number of PCIe Gen3/Gen2 slots available on POWER8 processor-based Power Systems.

2.8 External disk subsystems

This section describes the following external disk subsystems that can be attached to the Power 822 server:

- EXP24S SFF Gen2-bay drawer for high-density storage (#5887)
- IBM System Storage

Note: The EXP30 Ultra SSD Drawer (#EDR1 or #5888), the EXP12S SAS Disk Drawer (#5886), and the EXP24 SCSI Disk Drawer (#5786) are not supported on the Power S822 server.

2.8.1 EXP24S SFF Gen2-bay drawer

The EXP24S SFF Gen2-bay drawer (#5887) is an expansion drawer that supports up to 24 hot-swap 2.5-inch SFF SAS HDDs on POWER6, POWER6+, POWER7, POWER7+, or POWER8 servers in 2U of 19-inch rack space. The EXP24S drawer includes redundant AC power supplies and two power cords.

The EXP24S uses Gen2 or SFF-2 SAS drives that physically do not fit in the SFF-3 bays of the Power S822 system unit.

The EXP24S drawer is attached to SAS ports on either a PCIe SAS adapter located in the server or to the SAS ports on the rear of the server. Two SAS ports on the rear of the server are enabled with the expanded-function storage backplane with dual IOA support (#EJ0U).

The SAS controller and the EXP24S SAS ports are attached using the appropriate SAS Y or X cables.

The following internal SAS adapters support the EXP24S:

- PCIe2 LP RAID SAS Adapter Dual-port 6 Gb (#ESA2, CCIN 57B3)
- PCIe3 LP RAID SAS Adapter (#EJ0M, CCIN 57B4)

The SAS disk drives that are contained in the EXP24S SFF Gen2-bay Drawer are controlled by one or two PCIe SAS adapters that are connected to the EXP24S through SAS cables. The SAS cable varies, depending on the adapter being used, the operating system being used, and the protection you want.

In addition to the existing SAS disks options, IBM has the following available disk model:

- 1.2 TB 10K RPM SAS HDD in Gen-2 Carrier for AIX and Linux (#ESD3)

The EXP24S SFF Gen2-bay drawer can be ordered in one of three possible mode settings, configured by manufacturing (not customer set-up), of 1, 2, or 4 sets of disk bays.

With IBM AIX and Linux, the EXP24S can be ordered with four sets of six bays (mode 4), two sets of 12 bays (mode 2), or one set of 24 bays (mode 1).

There are six SAS connectors on the rear of the EXP24S drawer to which to SAS adapters or controllers are attached. They are labeled T1, T2, and T3; there are two T1, two T2, and two T3 connectors. Figure 2-24 shows the rear connectors of the EXP24S drawer.

- In mode 1, two or four of the six ports are used. Two T2 ports are used for a single SAS adapter, and two T2 and two T3 ports are used with a paired set of two adapters or dual adapters configuration.
- In mode 2 or mode 4, four ports are used, two T2 and two T3, to access all SAS bays.

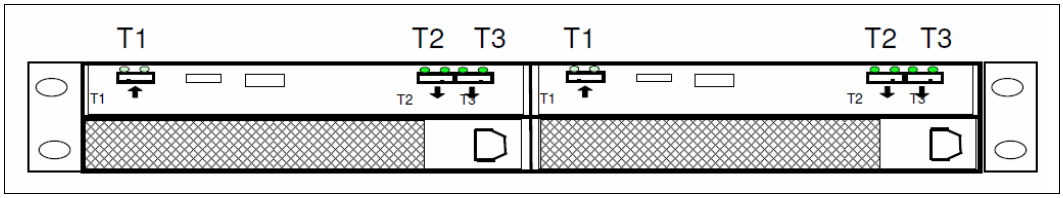


Figure 2-24 EXP24S SFF Gen2-bay drawer rear connectors

An EXP24S drawer in mode 4 can be attached to two or four SAS controllers and provide high configuration flexibility. An EXP24S in mode 2 has similar flexibility. Up to 24 HDDs can be supported with any of the supported SAS adapters or controllers.

Any EXP24S order includes the EXP24S drawer no-charge specify codes to indicate to IBM manufacturing the mode to which the drawer should be set. The drawer is delivered with this configuration. If multiple EXP24S drawers are ordered, mixing modes should not be within that order. There is no externally visible indicator regarding the drawer's mode.

Notes:

- The modes for the EXP24S drawer are set by IBM manufacturing. There is no option to reset after the drawer is shipped.
- One #5887 EXP24S drawer in mode 1 can be attached to the two SAS ports on the rear of the Power S822 server using two SAS YO cables such as #ECBT, #ECBU, #ECBV, or #ECBW cable options.
- Either SSDs or HDDs can be placed in this drawer, but SSDs and HDDs cannot be mixed in this drawer.
- Up to 14 EXP24S drawers can be attached to the Power S822 server.
- Longer distance SAS cables are thicker and can fill the cable management arm more quickly.

For details about the SAS cabling, see “SAS cabling for the 5887 drawer” in the information center:

http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/index.jsp?topic=/p7had/p7hadsascabling.htm

2.8.2 IBM System Storage

The IBM System Storage Disk Systems products and offerings provide compelling storage solutions with superior value for all levels of business, from entry-level to high-end storage systems. For more information a bout the various offerings, see the following website:

http://www.ibm.com/systems/storage/disk

The following section highlights a few of the offerings.

IBM Network Attached Storage (NAS)

IBM Network attached storage (NAS) products provide a wide-range of network attachment capabilities to a broad range of host and client systems, such as IBM Scale Out Network Attached Storage (SONAS) and the IBM System Storage Nxxx series. For more information about the hardware and software, see the following website:

http://www.ibm.com/systems/storage/network

IBM Storwize family

IBM Storwize® family is the ideal solution to optimize the data architecture for business flexibility and data storage efficiency. Different models such as the Storwize V3700, V5000, and V7000 offer Storage virtualization, Real-time Compression™, Easy Tier, and much more functionality. For more information, see the following website:

http://www.ibm.com/systems/storage/storwize

IBM Flash Storage

IBM flash storage delivers extreme performance to derive measurable economic value across the data architecture – servers, software, applications and storage. IBM offers a comprehensive flash portfolio with the IBM FlashSystem™ family. For more information, see the following website:

http://www.ibm.com/systems/storage/flash

IBM XIV Storage System

IBM XIV® is a high-end disk storage system, helping thousands of enterprises meet the challenge of data growth with hotspot-free performance and ease of use. Simple scaling, high service levels for dynamic, heterogeneous workloads, and tight integration with hypervisors and the OpenStack platform enable optimal storage agility for cloud environments.

XIV extends ease of use with integrated management for large and multi-site XIV deployments, reducing operational complexity and enhancing capacity planning. For more information, see the following website:

http://www.ibm.com/systems/storage/disk/xiv/index.html

IBM System Storage DS8000

The IBM System Storage DS8870 is a high-performance, high-capacity, secure storage system designed to deliver the highest levels of performance, flexibility, scalability, resiliency and total overall value for the most demanding, heterogeneous storage environments. The system is designed to manage a broad scope of storage workloads that exist in today's complex data center, doing it effectively and efficiently.

Additionally, the DS8000® includes a range of features that automate performance optimization and application quality of service, and also provide the highest levels of reliability and system uptime. For more information, see the following website:

http://www.ibm.com/systems/storage/disk/ds8000/index.html

2.9 Hardware Management Console (HMC)

The Hardware Management Console (HMC) is a dedicated appliance that allows administrators to configure and manage system resources on IBM Power Systems servers that use IBM POWER6, POWER6+ POWER7, POWER7+, and POWER8 processors. The HMC provides basic virtualization management support for configuring logical partitions (LPARs) and dynamic resource allocation, including processor and memory settings for selected Power Systems servers. The HMC also supports advanced service functions, including guided repair and verify, concurrent firmware updates for managed systems, and around-the-clock error reporting through IBM Electronic Service Agent™ for faster support.

The HMC management features help to improve server utilization, simplify systems management, and accelerate provisioning of server resources using the PowerVM virtualization technology.

Requirements: When using the HMC with the Power S822 servers, the HMC code must be running at V8R8.1.0 level, or later.

The Power S822 platforms support two main service environments:

- Attachment to one or more HMCs
  - This environment is the common configuration for servers supporting logical partitions with dedicated or virtual I/O. In this case, all servers have at least one logical partition.
- No HMC attachment

  Systems not attached to an HMC will fall under one of two service strategies:
  - Full system partition: A single partition owns all the server resources and only one operating system may be installed.
  - Partitioned system: The system has more than one partition, and the partitions can be running different operating systems. In this environment, partitions are managed by the Integrated Virtualization Manager (IVM), which includes some of the functions offered by the HMC.

Hardware support for customer-replaceable units comes standard along with the HMC. In addition, users have the option to upgrade this support level to IBM on-site support to be consistent with other Power Systems servers.

2.9.1 HMC code level

HMC V8R8.1.0 contains the following new features:

- Support for managing the IBM Power System S812L, S814, S822, S822L, and S824 servers
- Support for the new HMC model 7042-CR8

- PowerVM management enhancements:
  - Enhanced and simplified HMC management of PowerVM virtualization enables automation and simplifies the setup and operation of PowerVM
  - Improved “no-touch” VIOS virtualization management enables complete virtualization administration from the HMC
- Virtualization-oriented capacity and performance monitor and performance statistics

  These performance statistics can help you understand the workload characteristics and to plan for capacity
- HMC command to initiate a Remote Restart operation

  This removes the requirement of VMControl for the PowerVM Remote Restart function.
- Updated web browser support
  - Mozilla Firefox ESR 17 and 24
  - Microsoft Internet Explorer 9 through 11
  - Google Chrome 31

If you are attaching an HMC to a new server or adding function to an existing server that requires a firmware update, the HMC machine code may need to be updated to support the firmware level of the server. In a dual HMC configuration, both HMCs must be at the same version and release of the HMC code.

To determine the HMC machine code level required for the firmware level on any server, go to the following web page to access the Fix Level Recommendation Tool (FLRT) on or after the planned availability date for this product. FLRT will identify the correct HMC machine code for the selected system firmware level.

https://www14.software.ibm.com/webapp/set2/flrt/home

Note: Access to firmware and machine code updates is conditioned on entitlement and license validation in accordance with IBM policy and practice. IBM may verify entitlement through customer number, serial number electronic restrictions, or any other means or methods employed by IBM at its discretion.

2.9.2 HMC RAID 1 support

HMCs now offer a high-availability feature. The new 7042-CR8, by default, includes two hard drives with RAID 1 configured. RAID 1 is also offered on the 7042-CR6, 7042-CR7, and 7042-CR8 (if the feature was removed from the initial order) as an MES upgrade option.

RAID 1 uses data mirroring. Two physical drives are combined into an array, and the same data is written to both drives. This makes the drives a *mirror image* of each other. If one of the drives experiences a failure, it is taken off-line and the HMC continues operating with the other drive.

HMC models

To use an existing HMC to manage any POWER8 processor-based server, the HMC must be a model CR5, or later, rack-mounted HMC, or model C08, or later, deskside HMC. The latest HMC model is the 7042-CR8. For your reference, Table 2-20 on page 65 lists a comparison between the 7042-CR7 and the 7042-CR8 HMC models.

Table 2-20 Comparison between 7042-CR7 and 7042-CR8

Feature	CR7	CR8
IBM System x® model	x3550 M4	x3550 M4
HMC model	7042-CR7	7042-CR8
Processor	Intel Xeon E5	Intel Xeon E5
Memory	4 GB	16 GB
DASD	500 GB	500 GB
RAID 1	Default	Default
Multitech internal modem	Optional	None
USB ports	Two front, four back, one internal	Two front, four back, one internal
Integrated network	Four 1 Gb Ethernet	Four 1 Gb Ethernet
I/O slots	1 PCI Express 3.0 slot	1 PCI Express 3.0 slot

2.9.3 HMC connectivity to the POWER8 processor-based systems

POWER8 processor-based servers, and their predecessor systems, that are managed by an HMC require Ethernet connectivity between the HMC and the server's service processor. In addition, if dynamic LPAR, Live Partition Mobility, or PowerVM Active Memory Sharing operations are required on the managed partitions, Ethernet connectivity is needed between these partitions and the HMC. A minimum of two Ethernet ports are needed on the HMC to provide such connectivity.

For the HMC to communicate properly with the managed server, eth0 of the HMC must be connected to either the HMC1 or HMC2 ports of the managed server, although other network configurations are possible. You may attach a second HMC to the remaining HMC port of the server for redundancy. The two HMC ports must be addressed by two separate subnets. Figure 2-25 on page 66 shows a simple network configuration to enable the connection from the HMC to the server and to allow for dynamic LPAR operations. For more details about HMC and the possible network connections, see *IBM Power Systems HMC Implementation and Usage Guide*, SG24-7491.

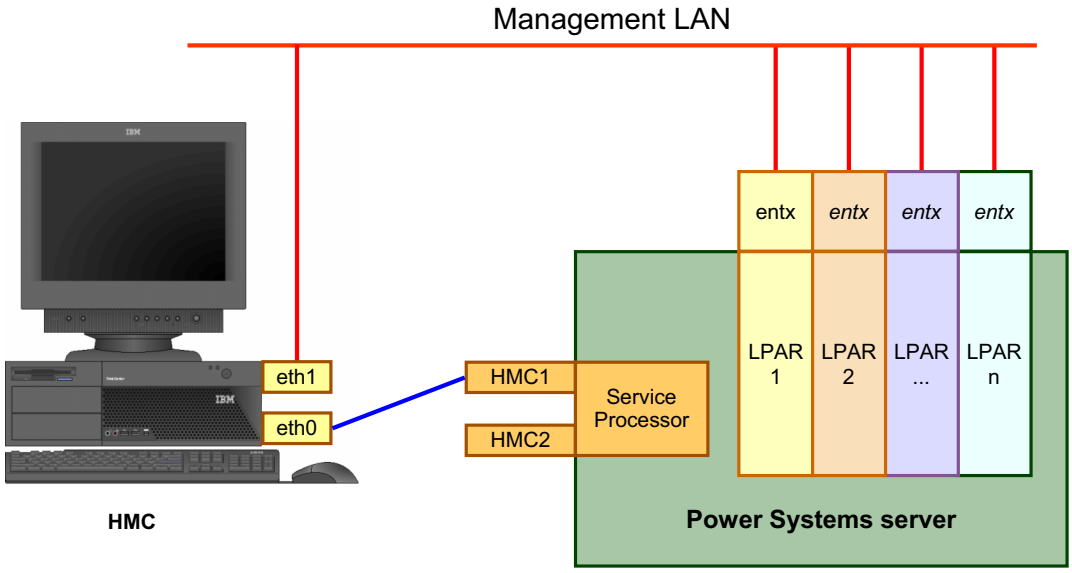


Figure 2-25 Network connections from HMC to service processor and LPARs

By default, the service processor HMC ports are configured for dynamic IP address allocation. The HMC can be configured as a DHCP server, providing an IP address at the time that the managed server is powered on. In this case, the flexible service processor (FSP) is allocated an IP address from a set of address ranges that are predefined in the HMC software.

If the service processor of the managed server does not receive a DHCP reply before time out, predefined IP addresses will be set up on both ports. Static IP address allocation is also an option and can be configured using the Advanced System Management Interface (ASMI) menus.

Notes: The two service processor HMC ports:

- run at a speed of 1 Gbps.
- are visible only to the service processor and can be used to attach the server to an HMC or to access the ASMI options from a client directly from a client web browser
- will use the following network configuration if no IP addresses are set:
  - service processor eth0 (HMC1 port): 169.254.2.147 with netmask 255.255.255.0
  - service processor eth1 (HMC2 port): 169.254.3.147 with netmask 255.255.255.0

For more information about the service processor, see “Service processor” on page 117.

2.9.4 High availability HMC configuration

The HMC is an important hardware component. While Power Systems servers and their hosted partitions can continue to operate when the managing HMC becomes unavailable, certain operations – such as dynamic LPAR, partition migration using PowerVM Live Partition Mobility, or the creation of a new partition – cannot be performed without the HMC. To avoid such situations, you may consider installing a second HMC, in a redundant configuration, to be available when the other is not – during maintenance, for example.

To achieve HMC redundancy for a POWER8 processor-based server, the server must be connected to two HMCs. The HMCs:

- must be running the same level of HMC code.
- must use different subnets to connect to the service processor.
- must be able to communicate with the server's partitions over a public network to allow for full synchronization and functionality.

Figure 2-26 shows one possible highly available HMC configuration that is managing two servers. Each HMC is connected to one FSP port of each managed server.

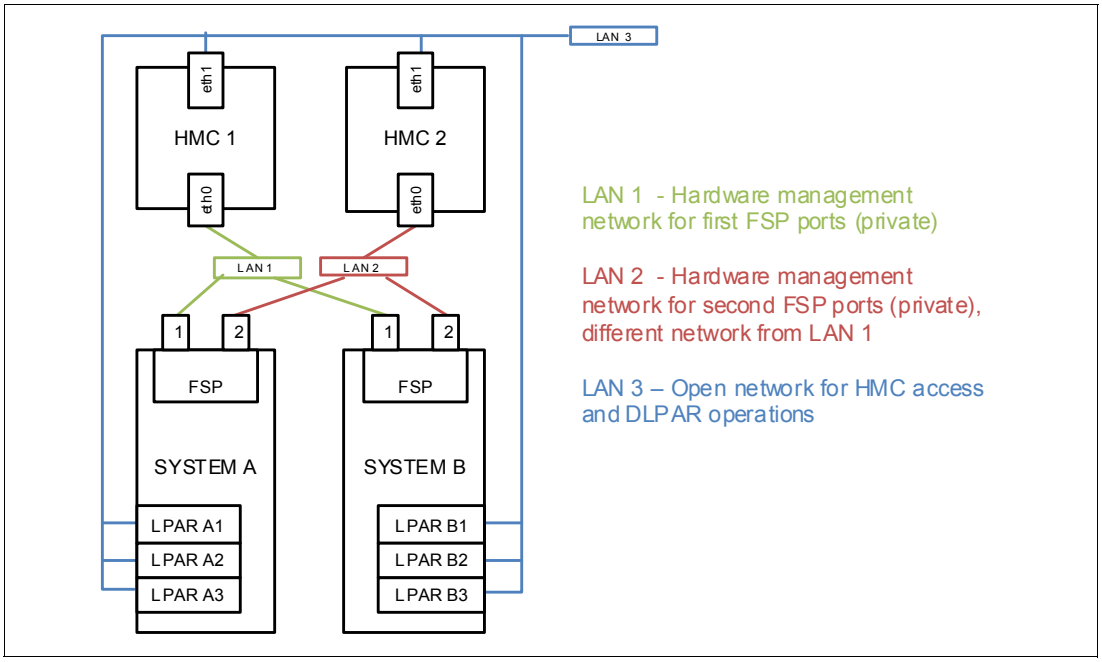


Figure 2-26 Highly available HMC networking example

For simplicity, only the hardware management networks (LAN1 and LAN2) are highly available (Figure 2-26). However, the open network (LAN3) can be made highly available by using a similar concept and adding a second network between the partitions and HMCs.

For details about redundant HMCs, see *IBM Power Systems HMC Implementation and Usage Guide*, SG24-7491.

2.10 Operating system support

The Power S822 servers support the following operating systems:

- AIX
- Linux

In addition, the Virtual I/O Server can be installed in special partitions that provide support to other partitions running AIX, or Linux operating systems for using features such as virtualized I/O devices, PowerVM Live Partition Mobility, or PowerVM Active Memory Sharing.

For details about the software available on IBM Power Systems, visit the IBM Power Systems Software™ website:

http://www.ibm.com/systems/power/software/index.html

2.10.1 AIX operating system

The following sections discuss the various levels of AIX operating system support.

IBM periodically releases maintenance packages (service packs or technology levels) for the AIX operating system. Information about these packages, downloading, and obtaining the CD-ROM is on the Fix Central website:

http://www-933.ibm.com/support/fixcentral/

The Fix Central website also provides information about how to obtain the fixes that are included on CD-ROM.

The Service Update Management Assistant (SUMA), which can help you to automate the task of checking and downloading operating system downloads, is part of the base operating system. For more information about the **suma** command, go to the following website:

http://www14.software.ibm.com/webapp/set2/sas/f/genunix/suma.html

AIX Version 6.1

The following minimum level of AIX Version 6.1 supports the Power S822:

- AIX Version 6.1 with the 6100-09 Technology Level and Service Pack 3, with APAR IV56366 or later

These additional AIX levels are supported in an LPAR using virtualized I/O only:

- AIX Version 6.1 with the 6100-09 Technology Level and Service Pack 1, or later
- AIX Version 6.1 with the 6100-08 Technology Level and Service Pack 1, or later
- AIX Version 6.1 with the 6100-07 Technology Level and Service Pack 6, or later

AIX Version 7.1

The following minimum level of AIX Version 7.1 supports the Power S822:

- AIX Version 7.1 with the 7100-03 Technology Level and Service Pack 3, with APAR IV56367 or later

These additional AIX levels are supported in an LPAR using virtualized I/O only:

- AIX Version 7.1 with the 7100-03 Technology Level and Service Pack 1, or later
- AIX Version 7.1 with the 7100-02 Technology Level and Service Pack 1, or later
- AIX Version 7.1 with the 7100-01 Technology Level and Service Pack 6, or later

Note:

- The POWER8 compatibility mode is supported on AIX Version 7.1 with the 7100-03 Technology Level and Service Pack 3 and later.
- All other prior AIX 7.1 levels and AIX 6.1 can run in POWER6, POWER6+, and POWER7 compatibility mode.
- The Easy Tier functionality that comes with the 18 bay storage backplane (#EJ0P) requires:
  - AIX Version 7.1 with the 7100-03 Technology Level and Service Pack 3 with APAR IV56367 or later
  - AIX Version 6.1 with the 6100-09 Technology Level and Service Pack 3 with APAR IV56366 or later

2.10.2 Linux operating system

Linux is an open source, cross-platform operating system that runs on numerous platforms from embedded systems to mainframe computers. It provides an implementation like UNIX across many computer architectures.

The supported versions of Linux on the Power S822 servers are as follows:

- Red Hat Enterprise Linux 6.5, or later
- SUSE Linux Enterprise Server 11 Service Pack 3, or later

Linux supports almost all of the Power System I/O and the configurator will verify support on order.

Be sure to connect your systems to the IBM Service and Productivity Tools for PowerLinux™ repository and keep up to date with the latest Linux for Power service and productivity tools, available from IBM:

http://www14.software.ibm.com/webapp/set2/sas/f/lopdiags/yum.html

For information about the PowerLinux Community refer to:

https://www.ibm.com/developerworks/group/tpl

For information about features and external devices that are supported by Linux, see the following site:

http://www.ibm.com/systems/power/software/linux/index.html

See more information about SUSE Linux Enterprise Server:

http://www.novell.com/products/server

See more information about Red Hat Enterprise Linux Advanced Server:

http://www.redhat.com/rhel/features

2.10.3 Virtual I/O Server

The minimum required level of Virtual I/O Server for both the Power S822 is VIOS 2.2.3.3.

Note: The Easy Tier functionality that comes with the eight bay storage backplane (#EJ0U) requires VIOS 2.2.3.3 with interim fix IV56366 or later

IBM regularly updates the Virtual I/O Server code. To find information about the latest updates, visit the Fix Central website:

http://www.ibm.com/support/fixcentral/

2.10.4 Java

There are unique considerations when running Java on POWER8 servers. For best use of the performance capabilities and most recent improvements of POWER8 technology, upgrade your JVM (or JDK) to IBM Java7 Release1 when possible, although IBM Java7, Java6, or Java5 still are able to run on POWER8. See the AIX Download and service information page:

http://www.ibm.com/developerworks/java/jdk/aix/service.html

2.11 Energy management

The Power S822 system is designed with features to help clients become more energy efficient. EnergyScale technology enables advanced energy management features to dramatically and dynamically conserve power and further improve energy efficiency. Intelligent Energy optimization capabilities enable the POWER8 processor to operate at a higher frequency for increased performance and performance per watt, or dramatically reduce frequency to save energy.

2.11.1 IBM EnergyScale technology

IBM EnergyScale technology provides functions to help the user understand and dynamically optimize processor performance versus processor energy consumption, and system workload, to control IBM Power Systems power and cooling usage.

EnergyScale uses power and thermal information that is collected from the system to implement policies that can lead to better performance or better energy utilization. IBM EnergyScale has the following features:

- Power trending

  EnergyScale provides continuous collection of real-time server energy consumption. It enables administrators to predict power consumption across their infrastructure and to react to business and processing needs. For example, administrators can use such information to predict data center energy consumption at various times of the day, week, or month.

- Power saver mode

  Power saver mode lowers the processor frequency and voltage on a fixed amount, reducing the energy consumption of the system while still delivering predictable performance. This percentage is predetermined to be within a safe operating limit and is not user configurable. The server is designed for a fixed frequency drop of almost 50% down from nominal frequency (the actual value depends on the server type and configuration).

Power saver mode is not supported during system startup, although it is a persistent condition that is sustained after the boot when the system starts executing instructions.

- Dynamic power saver mode

  Dynamic power saver mode varies processor frequency and voltage based on the utilization of the POWER8 processors. Processor frequency and utilization are inversely proportional for most workloads, implying that as the frequency of a processor increases, its utilization decreases, given a constant workload. Dynamic power saver mode takes advantage of this relationship to detect opportunities to save power, based on measured real-time system utilization.

  When a system is idle, the system firmware lowers the frequency and voltage to power energy saver mode values. When fully utilized, the maximum frequency varies, depending on whether the user favors power savings or system performance. If an administrator prefers energy savings and a system is fully utilized, the system is designed to reduce the maximum frequency to about 95% of nominal values. If performance is favored over energy consumption, the maximum frequency can be increased to up to 111.3% of nominal frequency for extra performance.

  Dynamic power saver mode is mutually exclusive with power saver mode. Only one of these modes can be enabled at a given time.

- Power capping

  Power capping enforces a user-specified limit on power usage. Power capping is not a power-saving mechanism. It enforces power caps by throttling the processors in the system, degrading performance significantly. The idea of a power cap is to set a limit that must never be reached but that frees extra power that was never used in the data center. The *margined* power is this amount of extra power that is allocated to a server during its installation in a data center. It is based on the server environmental specifications that usually are never reached because server specifications are always based on maximum configurations and worst-case scenarios.

- Soft power capping

  There are two power ranges into which the power cap can be set: power capping, as described previously, and soft power capping. Soft power capping extends the allowed energy capping range further, beyond a region that can be guaranteed in all configurations and conditions. If the energy management goal is to meet a particular consumption limit, then soft power capping is the mechanism to use.

- Processor core nap mode

  IBM POWER8 processor uses a low-power mode called *nap* that stops processor execution when there is no work to do on that processor core. The latency of exiting nap mode is small, typically not generating any impact on applications running. Therefore, the IBM POWER Hypervisor™ can use nap mode as a general-purpose idle state. When the operating system detects that a processor thread is idle, it yields control of a hardware thread to the POWER Hypervisor. The POWER Hypervisor immediately puts the thread into nap mode. Nap mode allows the hardware to turn the clock off on most of the circuits in the processor core. Reducing active energy consumption by turning off the clocks allows the temperature to fall, which further reduces leakage (static) power of the circuits causing a cumulative effect. Nap mode saves 10 - 15% of power consumption in the processor core.

- Processor core sleep mode

  To be able to save even more energy, the POWER8 processor has an even lower power mode referred to as *sleep*. Before a core and its associated private L2 cache enter sleep mode, the cache is flushed, transition lookaside buffers (TLB) are invalidated, and the hardware clock is turned off in the core and in the cache. Voltage is reduced to minimize

leakage current. Processor cores that are inactive in the system (such as capacity on demand, CoD, processor cores) are kept in sleep mode. Sleep mode saves about 80% power consumption in the processor core and its associated private L2 cache.

- Processor chip winkle mode

  The most amount of energy can be saved when a whole POWER8 chiplet enters the *winkle* mode. In this mode the entire chiplet is turned off including the L3 cache. This way can save more than 95% power consumption.

- Fan control and altitude input

  System firmware dynamically adjusts fan speed based on energy consumption, altitude, ambient temperature, and energy savings modes. Power Systems are designed to operate in worst-case environments, in hot ambient temperatures, at high altitudes, and with high-power components. In a typical case, one or more of these constraints are not valid. When no power savings setting is enabled, fan speed is based on ambient temperature and assumes a high-altitude environment. When a power savings setting is enforced (either Power Energy Saver Mode or Dynamic Power Saver Mode), fan speed will vary based on power consumption and ambient temperature.

- Processor folding

  Processor folding is a consolidation technique that dynamically adjusts, over the short term, the number of processors available for dispatch to match the number of processors demanded by the workload. As the workload increases, the number of processors made available increases. As the workload decreases, the number of processors that are made available decreases. Processor folding increases energy savings during periods of low to moderate workload because unavailable processors remain in low-power idle states (nap or sleep) longer.

- EnergyScale for I/O

  IBM POWER8 processor-based systems automatically power off hot pluggable PCI adapter slots that are empty or not being used. System firmware automatically scans all pluggable PCI slots at regular intervals, looking for those that meet the criteria for being not in use and powering them off. This support is available for all POWER8 processor-based servers and the expansion units that they support.

- Server power down

  If overall data center processor utilization is low, workloads can be consolidated on fewer numbers of servers so that some servers can be turned off completely. Consolidation makes sense when there will be long periods of low utilization, such as weekends. Live Partition Mobility can be used to move workloads to consolidate partitions onto fewer systems, reducing the number of servers powered on and therefore reducing the power usage.

On POWER8 processor-based systems, several EnergyScale technologies are embedded in the hardware and do not require an operating system or external management component. Fan control, environmental monitoring and system energy management are controlled by the On Chip Controller (OCC) and associated components. The power mode can also be set up without external tools, by using the ASMI interface, as shown in Figure 2-27.

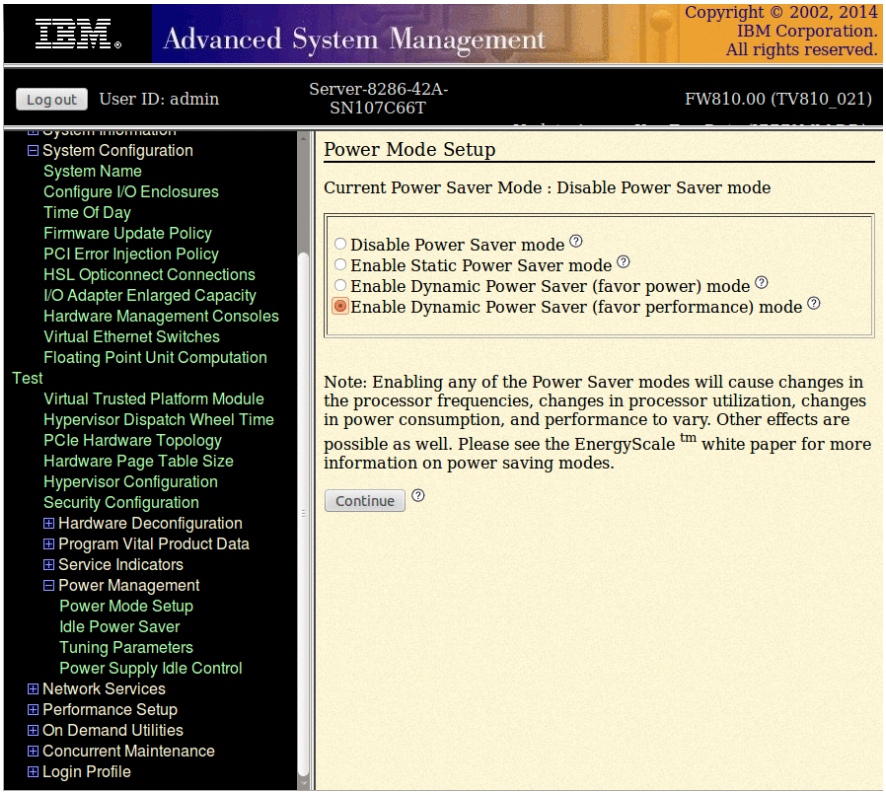

IBM
Advanced System Management
Copyright © 2002, 2014 IBM Corporation. All rights reserved.
Log out
User ID: admin
Server-8286-42A-SN107C66T
FW810.00 (TV810_021)
System Configuration
System Name
Configure I/O Enclosures
Time Of Day
Firmware Update Policy
PCI Error Injection Policy
HSL Opticonnect Connections
I/O Adapter Enlarged Capacity
Hardware Management Consoles
Virtual Ethernet Switches
Floating Point Unit Computation
Test
Virtual Trusted Platform Module
Hypervisor Dispatch Wheel Time
PCIe Hardware Topology
Hardware Page Table Size
Hypervisor Configuration
Security Configuration
Hardware Deconfiguration
Program Vital Product Data
Service Indicators
Power Management
Power Mode Setup
Idle Power Saver
Tuning Parameters
Power Supply Idle Control
Network Services
Performance Setup
On Demand Utilities
Concurrent Maintenance
Login Profile
Power Mode Setup
Current Power Saver Mode : Disable Power Saver mode
Disable Power Saver mode
Enable Static Power Saver mode
Enable Dynamic Power Saver (favor power) mode
Enable Dynamic Power Saver (favor performance) mode
Note: Enabling any of the Power Saver modes will cause changes in the processor frequencies, changes in processor utilization, changes in power consumption, and performance to vary. Other effects are possible as well. Please see the EnergyScale tm white paper for more information on power saving modes.
Continue

Figure 2-27 Setting the power mode in ASMI

2.11.2 On Chip Controller

To maintain the power dissipation of POWER7+ in spite of its large increase in performance and bandwidth, POWER8 invested significantly in power management innovations. A new on-chip controller (OCC) utilizing an embedded PowerPC® core with 512 KB of SRAM runs real-time control firmware to respond to workload variations by adjusting the per-core frequency and voltage based on activity, thermal, voltage, and current sensors.

The on-die nature of the OCC allows for approximately 100× speed up in response to workload changes over POWER7+, enabling reaction under the timescale of a typical OS timeslice and allowing for multi-socket, scalable systems to be supported. It also enables more granularity in controlling the energy parameters in the processor, and increases reliability in energy management by having one controller in each processor that can perform certain functions independently of the others.

POWER8 also includes an internal voltage regulation capability that enables each core to run at a different voltage. Optimizing both voltage and frequency for workload variation enables better increase in power savings versus optimizing frequency only.

2.11.3 Energy consumption estimation

Often, for Power Systems, various energy-related values are important:

- Maximum power consumption and power source loading values

  These values are important for site planning and are in the hardware information center:

  http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/index.jsp

  Search for type and model number and `server specifications`. For example, for the Power S822 system search for `8284-22A server specifications`.

- An estimation of the energy consumption for a certain configuration

  The calculation of the energy consumption for a certain configuration could be done in the IBM Systems Energy Estimator:

  http://www-912.ibm.com/see/EnergyEstimator/

  In that tool select the type and model for the desired system, enter some details of the configuration and a desired CPU utilization. As a result the tool shows the estimated energy consumption and the waste heat at the desired utilization and also at full utilization.

3

Virtualization

As you look for ways to maximize the return on your IT infrastructure investments, consolidating workloads becomes an attractive proposition.

IBM Power Systems combined with PowerVM technology offer key capabilities that can help you consolidate and simplify your IT environment:

- Improve server utilization and sharing I/O resources to reduce total cost of ownership and make better use of IT assets.
- Improve business responsiveness and operational speed by dynamically re-allocating resources to applications as needed, to better match changing business needs or handle unexpected changes in demand.
- Simplify IT infrastructure management by making workloads independent of hardware resources, so you can make business-driven policies to deliver resources based on time, cost, and service-level requirements.

This chapter discusses the virtualization technologies and features on IBM Power Systems:

- POWER Hypervisor
- POWER processor modes
- Active Memory Expansion
- PowerVM Simplification
- System Planning Tool
- PowerVC
- PowerVP™
- VIOS 2.2.3 features

© Copyright IBM Corp. 2014. All rights reserved.

3.1 POWER Hypervisor

Combined with features in the POWER8 processors, the IBM POWER Hypervisor delivers functions that enable other system technologies, including logical partitioning technology, virtualized processors, IEEE VLAN-compatible virtual switch, virtual SCSI adapters, virtual Fibre Channel adapters, and virtual consoles. The POWER Hypervisor is a basic component of the system's firmware and offers the following functions:

- Provides an abstraction between the physical hardware resources and the logical partitions that use them.
- Enforces partition integrity by providing a security layer between logical partitions.
- Controls the dispatch of virtual processors to physical processors (see "Processing mode" on page 88).
- Saves and restores all processor state information during a logical processor context switch.
- Controls hardware I/O interrupt management facilities for logical partitions.
- Provides virtual LAN channels between logical partitions that help to reduce the need for physical Ethernet adapters for inter-partition communication.
- Monitors the service processor and performs a reset or reload if it detects the loss of the service processor, notifying the operating system if the problem is not corrected.

The POWER Hypervisor is always active, regardless of the system configuration and also when not connected to the managed console. It requires memory to support the resource assignment to the logical partitions on the server. The amount of memory that is required by the POWER Hypervisor firmware varies according to several factors:

- Number of logical partitions
- Number of physical and virtual I/O devices used by the logical partitions
- Maximum memory values specified in the logical partition profiles

The minimum amount of physical memory that is required to create a partition will be the size of the system's logical memory block (LMB). The default LMB size varies according to the amount of memory that is configured in the CEC (Table 3-1).

Table 3-1 Configured CEC memory-to-default logical memory block size

Configurable CEC memory	Default logical memory block
Up to 32 GB	128 MB
Greater than 32 GB	256 MB

In most cases, however, the actual minimum requirements and recommendations of the supported operating systems are greater than 256 MB. Physical memory is assigned to partitions in increments of LMB.

The POWER Hypervisor provides the following types of virtual I/O adapters:

- Virtual SCSI
- Virtual Ethernet
- Virtual Fibre Channel
- Virtual (TTY) console

Virtual SCSI

The POWER Hypervisor provides a virtual SCSI mechanism for the virtualization of storage devices. The storage virtualization is accomplished by using two paired adapters:

- A virtual SCSI server adapter
- A virtual SCSI client adapter

A Virtual I/O Server partition can define virtual SCSI server adapters. Other partitions are *client* partitions. The Virtual I/O Server partition is a special logical partition, as described in 3.4.4, “Virtual I/O Server” on page 90. The Virtual I/O Server software is included on all PowerVM editions. When using the PowerVM Standard Edition and PowerVM Enterprise Edition, dual Virtual I/O Servers can be deployed to provide maximum availability for client partitions when performing Virtual I/O Server maintenance.

Virtual Ethernet

The POWER Hypervisor provides a virtual Ethernet switch function that allows partitions on the same server to use fast and secure communication without any need for physical interconnection. The virtual Ethernet allows a transmission speed up to 20 Gbps, depending on the maximum transmission unit (MTU) size, type of communication and CPU entitlement. Virtual Ethernet support began with IBM AIX Version 5.3, Red Hat Enterprise Linux 4, and SUSE Linux Enterprise Server, 9, and it is supported on all later versions. (For more information, see 3.4.8, “Operating system support for PowerVM” on page 96). The virtual Ethernet is part of the base system configuration.

Virtual Ethernet has the following major features:

- The virtual Ethernet adapters can be used for both IPv4 and IPv6 communication and can transmit packets with a size up to 65,408 bytes. Therefore, the maximum MTU for the corresponding interface can be up to 65,394 (or 65,390 if VLAN tagging is used).
- The POWER Hypervisor presents itself to partitions as a virtual 802.1Q-compliant switch. The maximum number of VLANs is 4096. Virtual Ethernet adapters can be configured as either untagged or tagged (following the IEEE 802.1Q VLAN standard).
- A partition can support 256 virtual Ethernet adapters. Besides a default port VLAN ID, the number of additional VLAN ID values that can be assigned per virtual Ethernet adapter is 20, which implies that each virtual Ethernet adapter can be used to access 21 virtual networks.
- Each partition operating system detects the virtual local area network (VLAN) switch as an Ethernet adapter without the physical link properties and asynchronous data transmit operations.

Any virtual Ethernet can also have connectivity outside of the server if a layer-2 bridge to a physical Ethernet adapter is set in one Virtual I/O Server partition, also known as Shared Ethernet Adapter. See 3.4.4, “Virtual I/O Server” on page 90, for details about shared Ethernet.

Adapter and access: Virtual Ethernet is based on the IEEE 802.1Q VLAN standard. No physical I/O adapter is required when creating a VLAN connection between partitions, and no access to an outside network is required.

Virtual Fibre Channel

A virtual Fibre Channel adapter is a virtual adapter that provides client logical partitions with a Fibre Channel connection to a storage area network through the Virtual I/O Server logical partition. The Virtual I/O Server logical partition provides the connection between the virtual Fibre Channel adapters on the Virtual I/O Server logical partition and the physical Fibre Channel adapters on the managed system. Figure 3-1 depicts the connections between the client partition virtual Fibre Channel adapters and the external storage. For additional information, see 3.4.8, “Operating system support for PowerVM” on page 96.

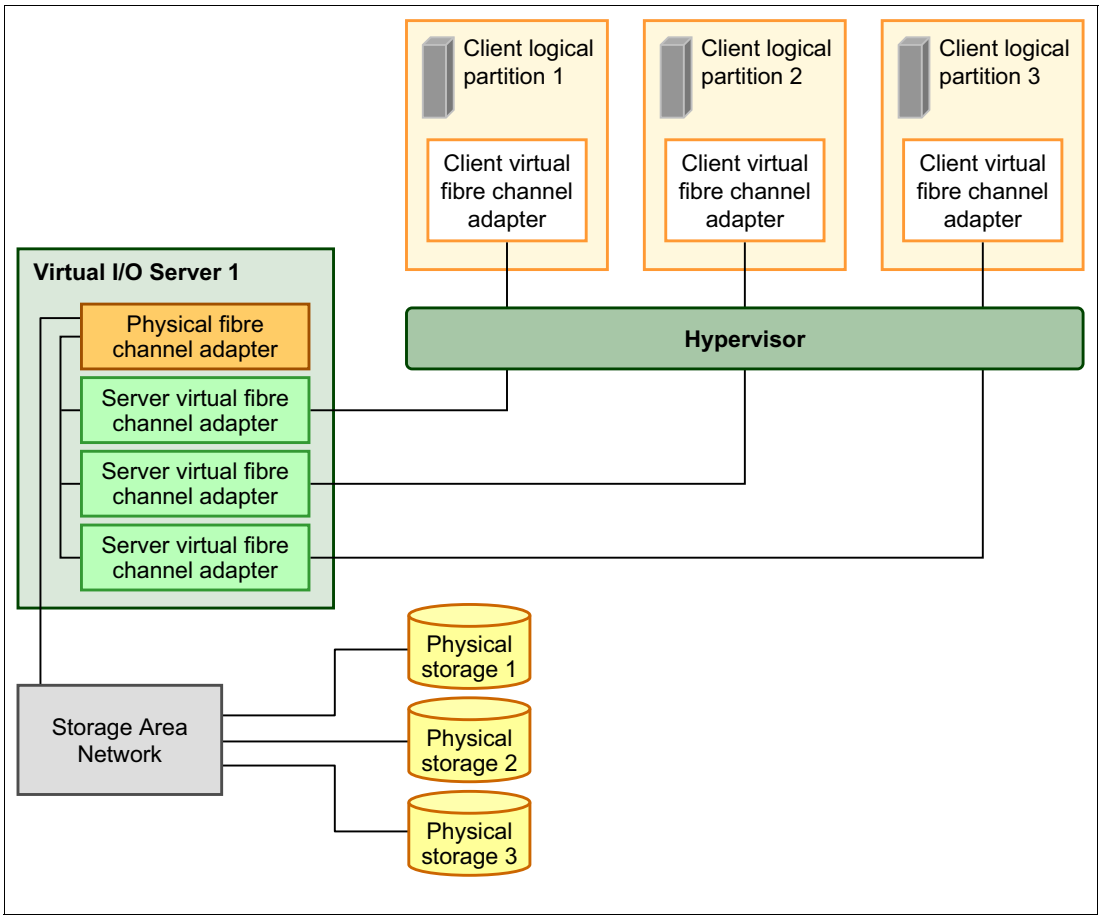


Figure 3-1 Connectivity between virtual Fibre Channels adapters and external SAN devices

Virtual (TTY) console

Each partition must have access to a system console. Tasks such as operating system installation, network setup, and various problem analysis activities require a dedicated system console. The POWER Hypervisor provides the virtual console by using a virtual TTY or serial adapter and a set of Hypervisor calls to operate on them. Virtual TTY does not require the purchase of any additional features or software, such as the PowerVM Edition features.

Depending on the system configuration, the operating system console can be provided by the Hardware Management Console virtual TTY, IVM virtual TTY, or from a terminal emulator that is connected to a system port.

3.2 POWER processor modes

Although, strictly speaking, not a virtualization feature, the POWER modes are described here because they affect various virtualization features.

On Power System servers, partitions can be configured to run in several modes, including the following modes:

- POWER6 compatibility mode

  This execution mode is compatible with Version 2.05 of the Power Instruction Set Architecture (ISA). For more information, visit the following address:

  http://power.org/wp-content/uploads/2012/07/PowerISA_V2.05.pdf

- POWER6+ compatibility mode

  This mode is similar to POWER6, with eight more storage protection keys.

- POWER7 mode

  This is the mode for POWER7+ and POWER7 processors, implementing the v2.06 of the Power Instruction Set Architecture. For more information, visit the following address:

  http://power.org/wp-content/uploads/2012/07/PowerISA_V2.06B_V2_PUBLIC.pdf

- POWER8 mode

  This is the native mode for POWER8 processors implementing the v2.07 of the Power Instruction Set Architecture. For more information, visit the following address:

  https://www.power.org/documentation/power-isa-version-2-07/

The selection of the mode is made on a per-partition basis, from the managed console, by editing the partition profile.

Figure 3-2 on page 80 shows the compatibility modes within the LPAR profile.

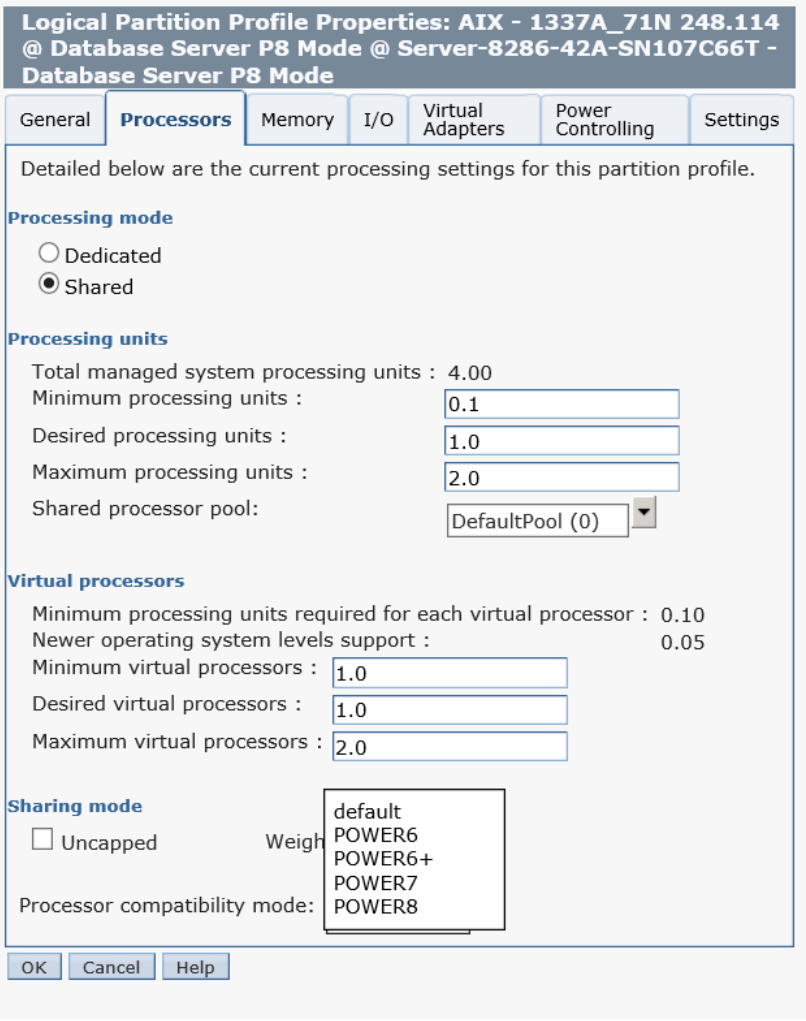


Figure 3-2 Configuring partition profile compatibility mode using the HMC

Table 3-2 lists the differences between processors modes.

Table 3-2 Differences between POWER6, POWER7, POWER8 compatibility mode

POWER6 and POWER6+ mode	POWER7 mode	POWER8 mode	Customer value
2-thread SMT	4-thread SMT	8-thread SMT	Throughput performance, processor core utilization
Vector Multimedia Extension/ AltiVec (VMX)	Vector scalar extension (VSX)	VSX2 In-Core Encryption Acceleration	High-performance computing
Affinity off by default	3-tier memory, micropartition affinity, dynamic platform optimizer	► HW memory affinity tracking assists ► Micropartition prefetch ► Concurrent LPARs per core	Improved system performance for system images spanning sockets and nodes
64-core and 128-thread scaling	► 32-core and 128-thread scaling ► 64-core and 256-thread scaling ► 128-core and 512-thread scaling ► 256-core and 1024-thread scaling	► 1024-thread Scaling ► Hybrid threads ► Transactional memory ► Active system optimization HW assists	Performance and scalability for large scale-up single system image workloads (such as OLTP, ERP scale-up, and WPAR consolidation)
EnergyScale CPU Idle	EnergyScale CPU Idle and Folding with NAP and SLEEP	WINKLE, NAP, SLEEP, Idle power saver	Improved energy efficiency

3.3 Active Memory Expansion

Active Memory Expansion is an optional feature for the Power S822 feature selecting the #4793 feature code in the e-Config tool.

This feature enables memory expansion on the system. By using compression and decompression of memory content can effectively expand the maximum memory capacity, providing additional server workload capacity and performance.

Active Memory Expansion is a technology that allows the effective maximum memory capacity to be much larger than the true physical memory maximum. Compression and decompression of memory content can allow memory expansion up to 125% for AIX partitions, which in turn enables a partition to perform significantly more work or support more users with the same physical amount of memory. Similarly, it can allow a server to run more partitions and do more work for the same physical amount of memory.

Note: The Active Memory Expansion feature is not supported with the Linux operating systems.

Active Memory Expansion uses the CPU resource of a partition to compress and decompress the memory contents of this same partition.The trade-off of memory capacity for processor cycles can be an excellent choice, but the degree of expansion varies based on how compressible the memory content is, and it also depends on having adequate spare CPU capacity available for this compression and decompression.

The POWER8 processor includes Active Memory Expansion on the processor chip to provide dramatic improvement in performance and greater processor efficiency. To take advantage of the hardware compression offload, AIX 6.1 Technology Level 8 is required.

Tests in IBM laboratories, using sample work loads, showed excellent results for many workloads in terms of memory expansion per additional CPU utilized. Other test workloads had more modest results. The ideal scenario is when there are many cold pages, that is, infrequently referenced pages. However, if many memory pages are referenced frequently, the Active Memory Expansion might not be a good choice.

Tip: If the workload is Java-based, the garbage collector must be tuned, so that it does not access the memory pages so often, turning cold pages to hot.

Clients have much control over Active Memory Expansion usage. Each individual AIX partition can turn on or turn off Active Memory Expansion. Control parameters set the amount of expansion you want in each partition to help control the amount of CPU that is used by the Active Memory Expansion function. An initial program load (IPL) is required for the specific partition that is turning memory expansion on or off. After turned on, monitoring capabilities are available in standard AIX performance tools, such as **lparstat**, **vmstat**, **topas**, and **svmon**. For specific POWER8 hardware compression, the **amepat** tool is used to configure the offload details.

Figure 3-3 represents the percentage of CPU that is used to compress memory for two partitions with separate profiles. Curve 1 corresponds to a partition that has spare processing power capacity. Curve 2 corresponds to a partition that is constrained in processing power.

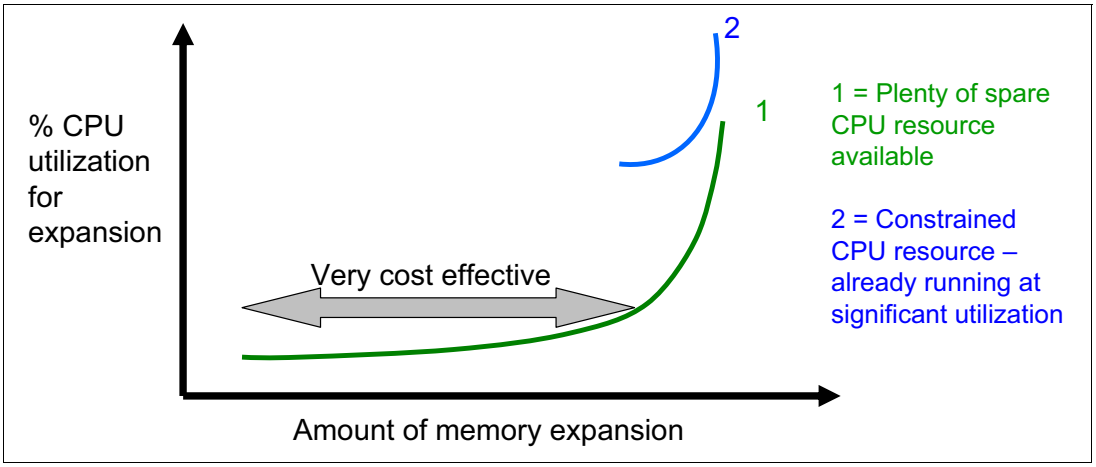

% CPU utilization for expansion
2
1
1 = Plenty of spare CPU resource available
2 = Constrained CPU resource – already running at significant utilization
Very cost effective
Amount of memory expansion

Figure 3-3 CPU usage versus memory expansion effectiveness

Both cases show that there is a “knee-of-curve” relationship for the CPU resource required for memory expansion:

- Busy processor cores do not have resources to spare for expansion.
- The more memory expansion is done, the more CPU resource is required.

The knee varies depending on how compressible the memory contents are. This example demonstrates the need for a case-by-case study of whether memory expansion can provide a positive return on investment.

To help you do this study, a the **amepat** planning tool is introduced with AIX 6.1 Technology Level 4 SP2, allowing you to sample actual workloads and estimate how expandable the partition's memory is and how much CPU resource is needed. Any model Power System can run the planning tool.

Figure 3-4 shows an example of the output that is returned by this planning tool. The tool outputs various real memory and CPU resource combinations to achieve the desired effective memory. It also recommends one particular combination. In this example, the tool recommends that you allocate 13% of processing power see Figure 3-4 to benefit from 119% extra memory capacity.

```
Active Memory Expansion Modeled Statistics:
-------------------------------------------
Modeled Expanded Memory Size : 52.00 GB
Achievable Compression ratio : 4.51

Expansion    Modeled True      Modeled              CPU Usage
Factor       Memory Size       Memory Gain          Estimate
---------    -------------     ------------------   -----------
    1.40         37.25 GB        14.75 GB [ 40%]   0.00 [  0%]
    1.80         29.00 GB        23.00 GB [ 79%]   0.87 [  5%]
    2.19         23.75 GB        28.25 GB [119%]   2.13 [ 13%]
    2.57         20.25 GB        31.75 GB [157%]   2.96 [ 18%]
    2.98         17.50 GB        34.50 GB [197%]   3.61 [ 23%]
    3.36         15.50 GB        36.50 GB [235%]   4.09 [ 26%]

Active Memory Expansion Recommendation:
---------------------------------------
The recommended AME configuration for this workload is to configure the LPAR
with a memory size of 23.75 GB and to configure a memory expansion factor
of 2.19.  This will result in a memory gain of 119%. With this
configuration, the estimated CPU usage due to AME is approximately 2.13
physical processors, and the estimated overall peak CPU resource required for
the LPAR is 11.65 physical processors.
```

Figure 3-4 Output from Active Memory Expansion planning tool

After you select the value of the memory expansion factor that you want to achieve, you can use this value to configure the partition from the managed console.

Figure 3-5 shows the activation of AME for each LPAR.

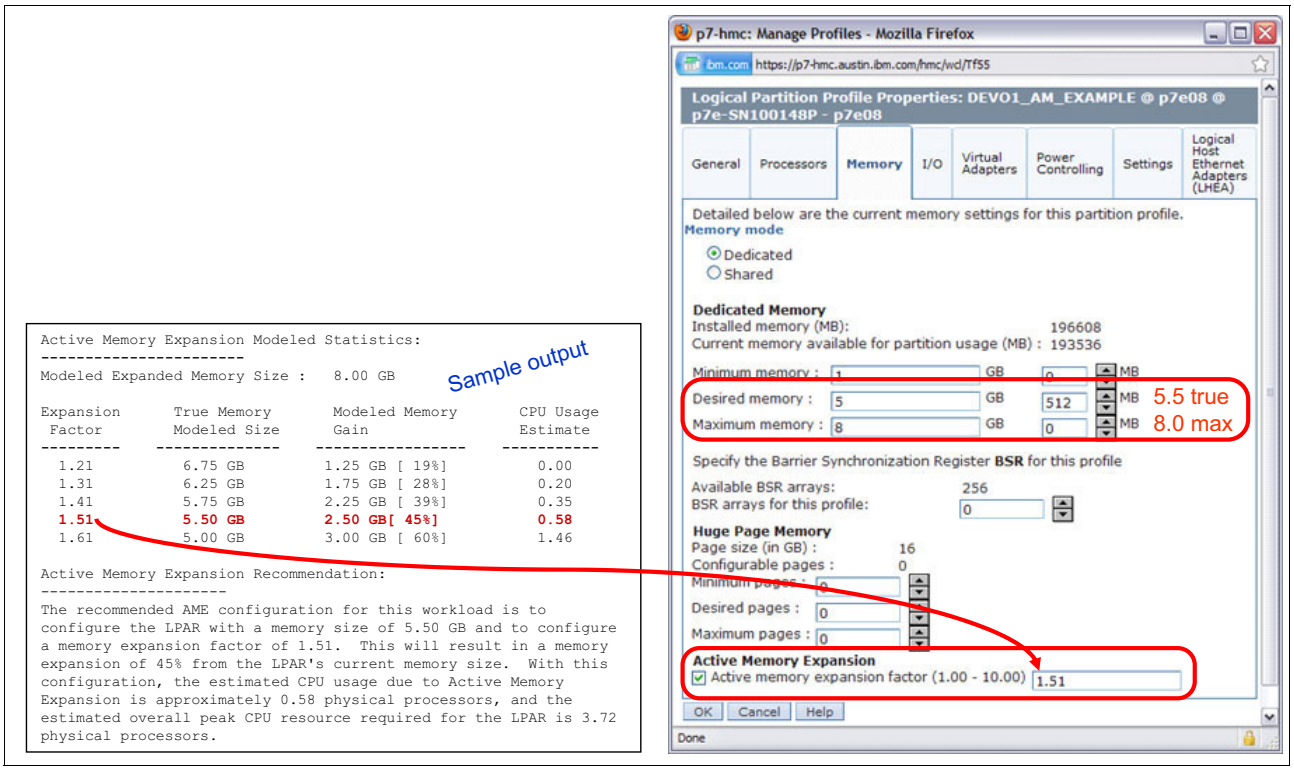

Active Memory Expansion Modeled Statistics:
Modeled Expanded Memory Size : 8.00 GB
Sample output
Expansion Factor | True Memory Modeled Size | Modeled Memory Gain | CPU Usage Estimate
1.21 6.75 GB 1.25 GB [19%] 0.00
1.31 6.25 GB 1.75 GB [28%] 0.20
1.41 5.75 GB 2.25 GB [39%] 0.35
1.51 5.50 GB 2.50 GB[45%] 0.58
1.61 5.00 GB 3.00 GB [60%] 1.46
Active Memory Expansion Recommendation:
The recommended AME configuration for this workload is to configure the LPAR with a memory size of 5.50 GB and to configure a memory expansion factor of 1.51. This will result in a memory expansion of 45% from the LPAR's current memory size. With this configuration, the estimated CPU usage due to Active Memory Expansion is approximately 0.58 physical processors, and the estimated overall peak CPU resource required for the LPAR is 3.72 physical processors.
p7-hmc: Manage Profiles - Mozilla Firefox
Logical Partition Profile Properties: DEVO1_AM_EXAMPLE @ p7e08 @ p7e-SN100148P - p7e08
General
Processors
Memory
I/O
Virtual Adapters
Power Controlling
Settings
Logical Host Ethernet Adapters (LHEA)
Detailed below are the current memory settings for this partition profile.
Memory mode
Dedicated
Shared
Dedicated Memory
Installed memory (MB): 196608
Current memory available for partition usage (MB) : 193536
Minimum memory :
Desired memory : 5 GB 512 MB
Maximum memory : 8 GB 0 MB
5.5 true
8.0 max
Specify the Barrier Synchronization Register BSR for this profile
Available BSR arrays: 256
BSR arrays for this profile: 0
Huge Page Memory
Page size (in GB) : 16
Configurable pages : 0
Minimum pages :
Desired pages : 0
Maximum pages : 0
Active Memory Expansion
Active memory expansion factor (1.00 - 10.00) 1.51
OK
Cancel
Help
Done

Figure 3-5 Using the planning tool result to configure the partition

On the HMC menu that describes the partition, select the **Active Memory Expansion** check box and enter the true and maximum memory, and the memory expansion factor. To turn off expansion, clear the check box. In both cases, reboot the partition to activate the change.

In addition, a one-time, 60-day trial of Active Memory Expansion is available to provide more exact memory expansion and CPU measurements. The trial can be requested by using the Power Systems Capacity on Demand web page:

http://www.ibm.com/systems/power/hardware/cod/

Active Memory Expansion can be ordered with the initial order of the server or as a miscellaneous equipment specification (MES) order. A software key is provided when the enablement feature is ordered that is applied to the server. Rebooting is not required to enable the physical server. The key is specific to an individual server and is permanent. It cannot be moved to a separate server. This feature is ordered per server, independent of the number of partitions using memory expansion.

From the HMC, you can view whether the Active Memory Expansion feature was activated for the server.

Figure 3-6 shows the available server capabilities.

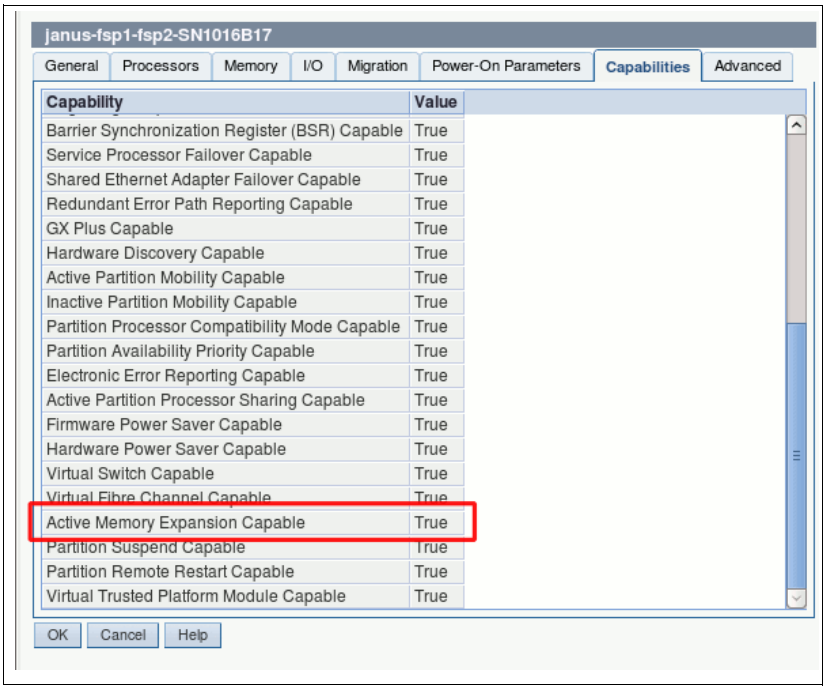


Figure 3-6 Server capabilities listed from the HMC

Moving an LPAR: If you want to move an LPAR that has Active Memory Expansion enabled to another physical server using Live Partition Mobility, the target server must have Active Memory Expansion activated with the software key). If the target system does not have Active Memory Expansion activated, the mobility operation fails during the premobility check phase, and an appropriate error message is displayed.

For details about Active Memory Expansion, download the document *Active Memory Expansion: Overview and Usage Guide*:

http://public.dhe.ibm.com/common/ssi/ecm/en/pow03037usen/POW03037USEN.PDF

3.4 PowerVM

The PowerVM platform is the family of technologies, capabilities, and offerings that delivers industry-leading virtualization on the IBM Power Systems. It is the umbrella branding term for Power Systems virtualization (Logical Partitioning, IBM Micro-Partitioning®, POWER Hypervisor, Virtual I/O Server, Live Partition Mobility, and more). As with Advanced Power Virtualization in the past, PowerVM is a combination of hardware enablement and value-added software. The licensed features of each of the three separate editions of PowerVM are described in 3.4.1, “PowerVM editions” on page 86.

3.4.1 PowerVM editions

The two editions of PowerVM are suited for various purposes:

- PowerVM Standard Edition

  This edition provides advanced virtualization functions and is intended for production deployments and server consolidation.

- PowerVM Enterprise Edition

  This edition is suitable for large server deployments such as multi-server deployments and cloud infrastructures. It includes unique features such as Active Memory Sharing and Live Partition Mobility.

Table 3-3 lists the feature codes of the PowerVM Editions that are available on the Power S822 server.

Table 3-3 Available PowerVM editions

Server	PowerVM Standard Edition	PowerVM Enterprise Edition
IBM Power S822	#5227	#5228

3.4.2 Logical partitions

Logical partitions (LPARs) and virtualization increase use of system resources and add a new level of configuration possibilities.

Logical partitioning

Logical partitioning was introduced with the POWER4 processor-based product line and the AIX Version 5.1, Red Hat Enterprise Linux 3.0 and SUSE Linux Enterprise Server 9.0 operating systems. This technology offered the capability to divide a system into separate logical systems, allowing each LPAR to run an operating environment on dedicated attached devices, such as processors, memory, and I/O components.

Later, dynamic logical partitioning increased the flexibility, allowing selected system resources, such as processors, memory, and I/O components, to be added and deleted from logical partitions while they are executing. AIX Version 5.2, with all the necessary enhancements to enable dynamic LPAR, was introduced in 2002. At the same time, Red Hat Enterprise Linux 5 and SUSE Linux Enterprise 9.0 were also able to support dynamic logical partitioning. The ability to reconfigure dynamic LPARs encourages system administrators to dynamically redefine all available system resources to reach the optimum capacity for each defined dynamic LPAR.

Micro-Partitioning

The IBM Micro-Partitioning technology allows you to allocate fractions of processors to a logical partition. This technology was introduced with POWER5 processor-based systems. A logical partition using fractions of processors is also known as a *shared processor partition* or micropartition. Micropartitions run over a set of processors called a *shared processor pool*, and virtual processors are used to let the operating system manage the fractions of processing power assigned to the logical partition. From an operating system perspective, a virtual processor cannot be distinguished from a physical processor, unless the operating system has been enhanced to be made aware of the difference. Physical processor are abstracted into virtual processors that are available to partitions. The meaning of the term *physical processor* in this section is a *processor core*.

When defining a shared processor partition, several options must be defined:

- The minimum, desired, and maximum processing units

  Processing units are defined as processing power, or the fraction of time that the partition is dispatched on physical processors. Processing units define the capacity entitlement of the partition.

- The shared processor pool

  Select one from the list with the names of each configured shared processor pool. This list also displays, in parentheses, the pool ID of each configured shared processor pool. If the name of the desired shared processor pool is not available here, you must first configure the shared processor pool by using the shared processor pool Management window. Shared processor partitions use the default shared processor pool, called DefaultPool by default. See 3.4.3, “Multiple shared processor pools” on page 89, for details about multiple shared processor pools.

- Whether the partition will be able to access extra processing power to “fill up” its virtual processors above its capacity entitlement (selecting either to cap or uncap your partition).

  If spare processing power is available in the shared processor pool or other partitions are not using their entitlement, an uncapped partition can use additional processing units if its entitlement is not enough to satisfy its application processing demand.

- The weight (preference) in the case of an uncapped partition.

- The minimum, desired, and maximum number of virtual processors.

The POWER Hypervisor calculates partition processing power based on minimum, desired, and maximum values, processing mode, and is also based on requirements of other active partitions. The actual entitlement is never smaller than the processing unit’s desired value, but can exceed that value in the case of an uncapped partition and up to the number of virtual processors allocated.

On the POWER8 processors, a partition can be defined with a processor capacity as small as 0.05 processing units. This number represents 0.05 of a physical core. Each physical core can be shared by up to 20 shared processor partitions, and the partition’s entitlement can be incremented fractionally by as little as 0.01 of the processor. The shared processor partitions are dispatched and time-sliced on the physical processors under control of the POWER Hypervisor. The shared processor partitions are created and managed by the HMC.

The Power S822 supports up to 20 cores in a single system, and these maximum numbers:

- 20 dedicated partitions
- 400 micropartitions (maximum 20 micropartitions per physical active core)

An important point is that the maximum amounts are supported by the hardware, but the practical limits depend on application workload demands.

Consider the following additional information about virtual processors:

- A virtual processor can be running (dispatched) either on a physical core or as standby waiting for a physical core to became available.

- Virtual processors do not introduce any additional abstraction level. They are only a dispatch entity. When running on a physical processor, virtual processors run at the same speed as the physical processor.

- Each partition’s profile defines CPU entitlement that determines how much processing power any given partition should receive. The total sum of CPU entitlement of all partitions cannot exceed the number of available physical processors in a shared processor pool.

- The number of virtual processors can be changed dynamically through a dynamic LPAR operation.

Processing mode

When you create a logical partition, you can assign entire processors for dedicated use, or you can assign partial processing units from a shared processor pool. This setting defines the processing mode of the logical partition. Figure 3-7 shows a diagram of the concepts described in this section.

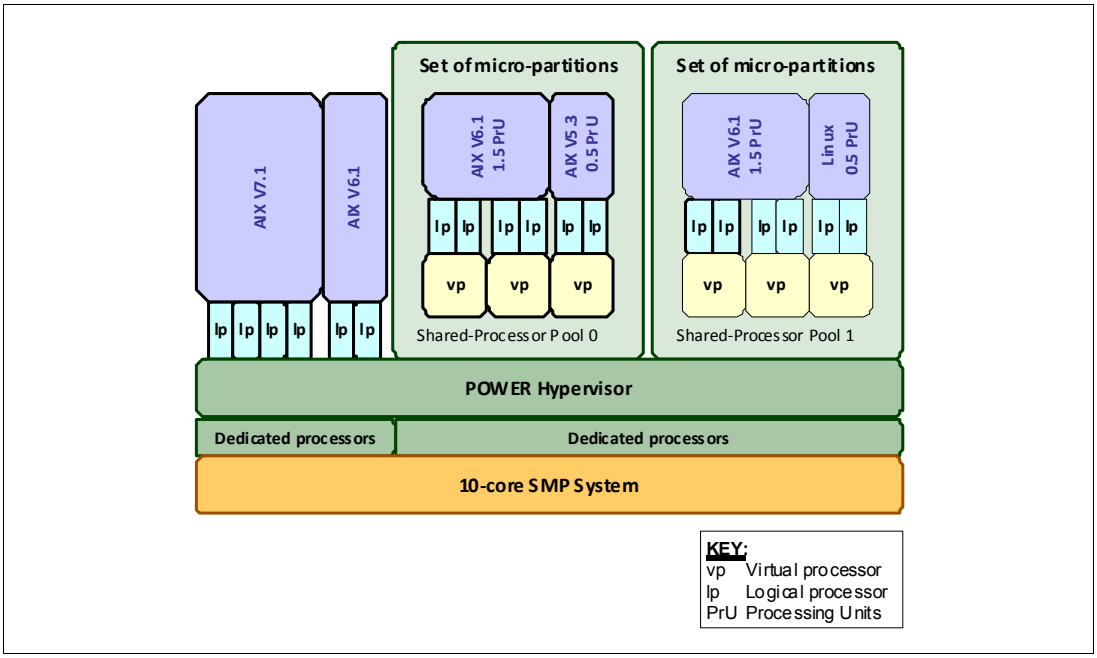


Figure 3-7 Logical partitioning concepts

Dedicated mode

In dedicated mode, physical processors are assigned as a whole to partitions. The simultaneous multithreading feature in the POWER8 processor core allows the core to execute instructions from two, four or eight independent software threads simultaneously. To support this feature consider the concept of *logical processors*. The operating system (AIX or Linux) sees one physical core as two, four or eight logical processors if the simultaneous multithreading feature is on. It can be turned off and on dynamically while the operating system is executing (for AIX, use the **smtctl** command; for Linux, use the **ppc64_cpu --smt** command). If simultaneous multithreading is off, each physical core is presented as one logical processor in AIX or Linux, and thus only one thread.

Shared dedicated mode

On POWER8 processor technology-based servers, you can configure dedicated partitions to become processor donors for idle processors that they own, allowing for the donation of spare CPU cycles from dedicated processor partitions to a shared processor pool. The dedicated partition maintains absolute priority for dedicated CPU cycles. Enabling this feature can help to increase system utilization without compromising the computing power for critical workloads in a dedicated processor.

Shared mode

In shared mode, logical partitions use virtual processors to access fractions of physical processors. Shared partitions can define any number of virtual processors (the maximum

number is 20 times the number of processing units that are assigned to the partition). From the POWER Hypervisor perspective, virtual processors represent dispatching objects. The POWER Hypervisor dispatches virtual processors to physical processors according to the partition's processing units entitlement. One processing unit represents one physical processor's processing capacity. At the end of the POWER Hypervisor's dispatch cycle (10 ms), all partitions receive total CPU time equal to their processing unit's entitlement. The logical processors are defined on top of virtual processors. So, even with a virtual processor, the concept of a logical processor exists and the number of logical processors depends whether the simultaneous multithreading is turned on or off.

3.4.3 Multiple shared processor pools

Multiple shared processor pools (MSPPs) is a capability that is supported on POWER8 processor-based servers. This capability allows a system administrator to create a set of micropartitions with the purpose of controlling the processor capacity that can be consumed from the physical shared processor pool.

Implementing MSPPs depends on a set of underlying techniques and technologies. Figure 3-8 is an overview of the architecture of multiple shared processor pools.

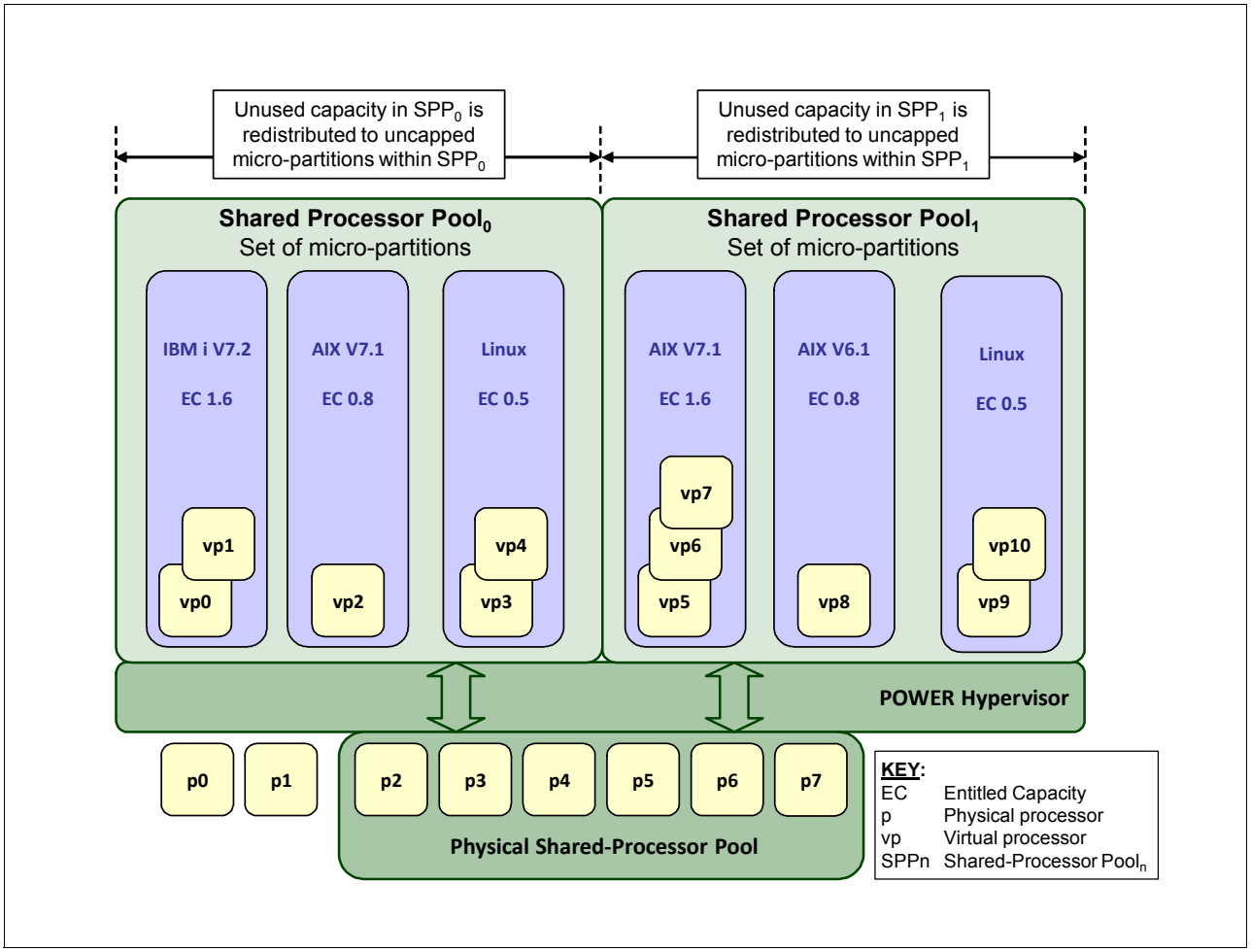


Figure 3-8 Overview of the architecture of multiple shared processor pools

Micropartitions are created and then identified as members of either the default shared processor pool$_0$ or a user-defined shared processor pool$_n$. The virtual processors that exist

within the set of micropartitions are monitored by the POWER Hypervisor, and processor capacity is managed according to user-defined attributes.

If the Power Systems server is under heavy load, each micropartition within a shared processor pool is guaranteed its processor entitlement plus any capacity that it might be allocated from the reserved pool capacity if the micropartition is uncapped.

If certain micropartitions in a shared processor pool do not use their capacity entitlement, the unused capacity is ceded and other uncapped micropartitions within the same shared processor pool are allocated the additional capacity according to their uncapped weighting. In this way, the entitled pool capacity of a shared processor pool is distributed to the set of micropartitions within that shared processor pool.

All Power Systems servers that support the multiple shared processor pools capability have a minimum of one (the default) shared processor pool and up to a maximum of 64 shared processor pools.

For more information and requirements about the shared storage pool, see the following publication: *IBM PowerVM Virtualization Introduction and Configuration,* SG24-7490.

3.4.4 Virtual I/O Server

The Virtual I/O Server is part of all PowerVM editions. It is a special-purpose partition that allows the sharing of physical resources between logical partitions to allow more efficient utilization (for example, consolidation). In this case, the Virtual I/O Server owns the physical resources (SCSI, Fibre Channel, network adapters, and optical devices) and allows client partitions to share access to them, thus minimizing the number of physical adapters in the system. The Virtual I/O Server eliminates the requirement that every partition owns a dedicated network adapter, disk adapter, and disk drive. The Virtual I/O Server supports OpenSSH for secure remote logins. It also provides a firewall for limiting access by ports, network services, and IP addresses. Figure 3-9 shows an overview of a Virtual I/O Server configuration.

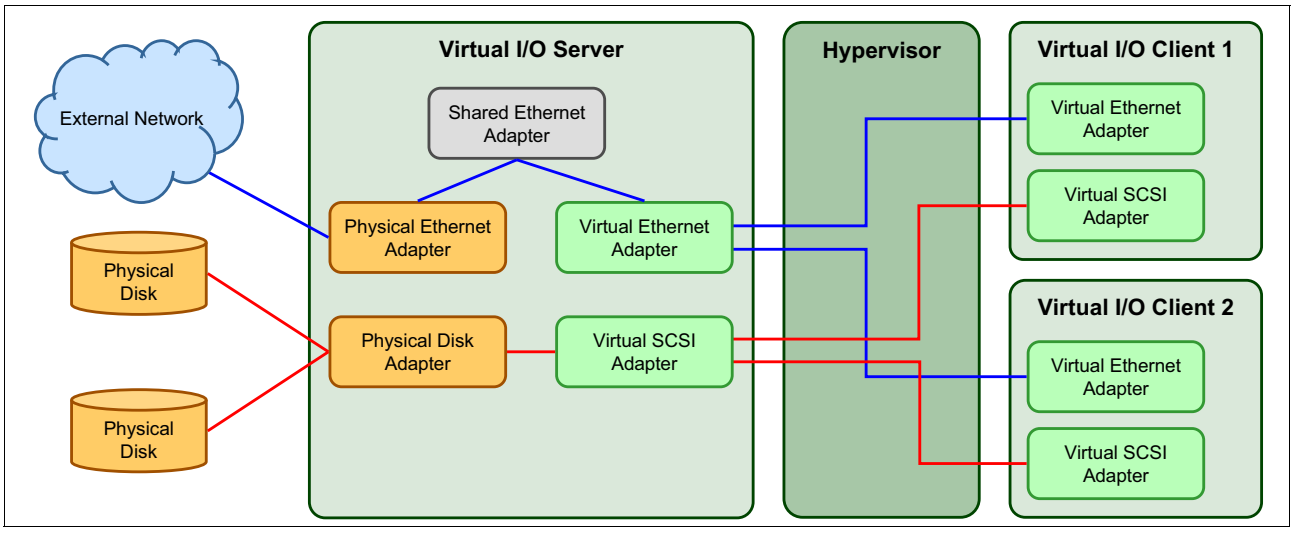


Figure 3-9 Architectural view of the Virtual I/O Server

Because the Virtual I/O Server is an operating system-based appliance server, redundancy for physical devices attached to the Virtual I/O Server can be provided by using capabilities such as Multipath I/O and IEEE 802.3ad Link Aggregation.

Installation of the Virtual I/O Server partition is performed from a special system backup DVD that is provided to clients who order any PowerVM edition. This dedicated software is only for the Virtual I/O Server, and is supported only in special Virtual I/O Server partitions. Three major virtual devices are supported by the Virtual I/O Server:

- Shared Ethernet Adapter
- Virtual SCSI
- Virtual Fibre Channel adapter

The Virtual Fibre Channel adapter is used with the NPIV feature, described in 3.4.8, “Operating system support for PowerVM” on page 96.

Shared Ethernet Adapter

A Shared Ethernet Adapter (SEA) can be used to connect a physical Ethernet network to a virtual Ethernet network. The Shared Ethernet Adapter provides this access by connecting the POWER Hypervisor VLANs with the VLANs on the external switches. Because the Shared Ethernet Adapter processes packets at layer 2, the original MAC address and VLAN tags of the packet are visible to other systems on the physical network. IEEE 802.1 VLAN tagging is supported.

The Shared Ethernet Adapter also provides the ability for several client partitions to share one physical adapter. With an SEA, you can connect internal and external VLANs by using a physical adapter. The Shared Ethernet Adapter service can be hosted only in the Virtual I/O Server, not in a general-purpose AIX or Linux partition, and acts as a layer-2 network bridge to securely transport network traffic between virtual Ethernet networks (internal) and one or more (EtherChannel) physical network adapters (external). These virtual Ethernet network adapters are defined by the POWER Hypervisor on the Virtual I/O Server.

Figure 3-10 shows a configuration example of an SEA with one physical and two virtual Ethernet adapters. An SEA can include up to 16 virtual Ethernet adapters on the Virtual I/O Server that share the same physical access.

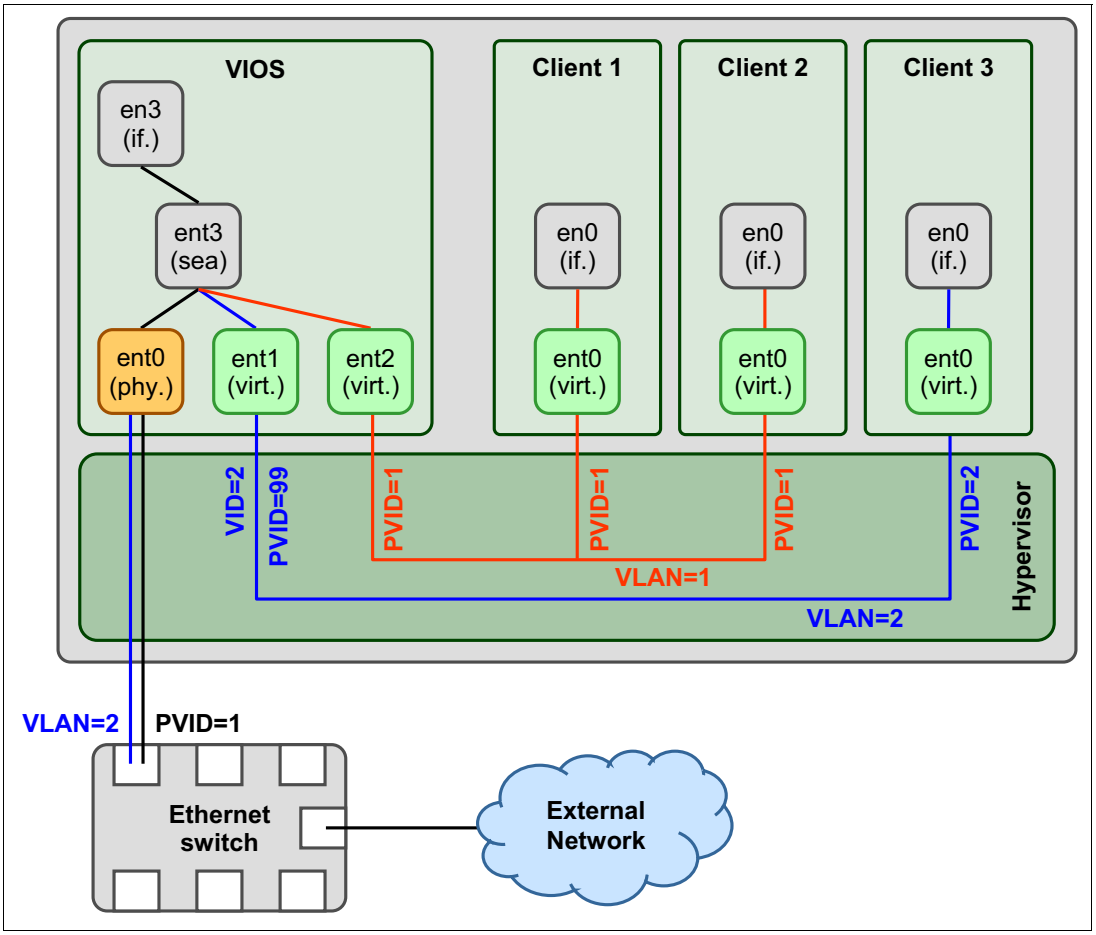


Figure 3-10 Architectural view of a Shared Ethernet Adapter

A single SEA setup can have up to 16 virtual Ethernet trunk adapters and each virtual Ethernet trunk adapter can support up to 20 VLAN networks. Therefore, a possibility is for a single physical Ethernet to be shared between 320 internal VLAN networks. The number of shared Ethernet adapters that can be set up in a Virtual I/O Server partition is limited only by the resource availability, because there are no configuration limits.

Unicast, broadcast, and multicast are supported, so protocols that rely on broadcast or multicast, such as Address Resolution Protocol (ARP), Dynamic Host Configuration Protocol (DHCP), Boot Protocol (BOOTP), and Neighbor Discovery Protocol (NDP), can work on an SEA.

Virtual SCSI

Virtual SCSI is used to see a virtualized implementation of the SCSI protocol. Virtual SCSI is based on a client/server relationship. The Virtual I/O Server logical partition owns the physical resources and acts as a server or, in SCSI terms, a target device. The client logical partitions access the virtual SCSI backing storage devices provided by the Virtual I/O Server as clients.

The virtual I/O adapters (virtual SCSI server adapter and a virtual SCSI client adapter) are configured using a managed console or through the Integrated Virtualization Manager on

smaller systems. The virtual SCSI server (target) adapter is responsible for executing any SCSI commands that it receives. It is owned by the Virtual I/O Server partition. The virtual SCSI client adapter allows a client partition to access physical SCSI and SAN attached devices and LUNs that are assigned to the client partition. The provisioning of virtual disk resources is provided by the Virtual I/O Server.

Physical disks that are presented to the Virtual/O Server can be exported and assigned to a client partition in various ways:

- The entire disk is presented to the client partition.
- The disk is divided into several logical volumes, which can be presented to a single client or multiple clients.
- As of Virtual I/O Server 1.5, files can be created on these disks, and file-backed storage devices can be created.

The logical volumes or files can be assigned to separate partitions. Therefore, virtual SCSI enables sharing of adapters and disk devices.

For more information about specific storage devices that are supported for Virtual I/O Server, see the following web page:

http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/datasheet.html

N_Port ID Virtualization

N_Port ID Virtualization (NPIV) is a technology that allows multiple logical partitions to access independent physical storage through the same physical Fibre Channel adapter. This adapter is attached to a Virtual I/O Server partition that acts only as a pass-through, managing the data transfer through the POWER Hypervisor.

Each partition that uses NPIV is identified by a pair of unique worldwide port names, enabling you to connect each partition to independent physical storage on a SAN. Unlike virtual SCSI, only the client partitions see the disk.

For more information and requirements for NPIV, see the following resources:

- *PowerVM Migration from Physical to Virtual Storage*, SG24-7825
- *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590

Virtual I/O Server functions

The Virtual I/O Server has many features, including monitoring solutions:

- Support for Live Partition Mobility starting on POWER6 processor-based systems with the PowerVM Enterprise Edition. For more information about Live Partition Mobility, see 3.4.5, “PowerVM Live Partition Mobility” on page 94.
- Support for virtual SCSI devices backed by a file, which are then accessed as standard SCSI-compliant LUNs.
- Support for virtual Fibre Channel devices that are used with the NPIV feature.
- Virtual I/O Server Expansion Pack with additional security functions such as Kerberos (Network Authentication Service for users and client and server applications), Simple Network Management Protocol (SNMP) v3, and Lightweight Directory Access Protocol (LDAP) client functionality.
- System Planning Tool (SPT) and Workload Estimator, which are designed to ease the deployment of a virtualized infrastructure. For more information about the System Planning Tool, see 3.5, “System Planning Tool” on page 101.

- IBM Systems Director agent and several preinstalled IBM Tivoli® agents, such as the following examples:
  - Tivoli Identity Manager, to allow easy integration into an existing Tivoli Systems Management infrastructure
  - Tivoli Application Dependency Discovery Manager (ADDM), which creates and automatically maintains application infrastructure maps including dependencies, change-histories, and deep configuration values
- vSCSI enterprise reliability, availability, serviceability (eRAS).
- Additional CLI statistics in **svmon**, **vmstat**, **fcstat**, and **topas**.
- VIOS Performance Advisor tool provides advisory reports based on key performance metrics for various partition resources collected from the VIOS environment
- Monitoring solutions to help manage and monitor the Virtual I/O Server and shared resources. Commands and views provide additional metrics for memory, paging, processes, Fibre Channel HBA statistics, and virtualization.

For more information about the Virtual I/O Server and its implementation, see: *IBM PowerVM Virtualization Introduction and Configuration,* SG24-7940

3.4.5 PowerVM Live Partition Mobility

PowerVM Live Partition Mobility allows you to move a running logical partition, including its operating system and running applications, from one system to another without any shutdown or without disrupting the operation of that logical partition. Inactive partition mobility allows you to move a powered-off logical partition from one system to another.

Live Partition Mobility provides systems management flexibility and improves system availability:

- Avoid planned outages for hardware or firmware maintenance by moving logical partitions to another server and then performing the maintenance. Live Partition Mobility can help lead to zero downtime maintenance because you can use it to work around scheduled maintenance activities.
- Avoid downtime for a server upgrade by moving logical partitions to another server and then performing the upgrade. This approach allows your users to continue their work without disruption.
- Avoid unplanned downtime. With preventive failure management, if a server indicates a potential failure, you can move its logical partitions to another server before the failure occurs. Partition mobility can help avoid unplanned downtime.
- Take advantage of server optimization:
  - Consolidation: You can consolidate workloads that run on several small, under-used servers onto a single large server.
  - Deconsolidation: You can move workloads from server to server to optimize resource use and workload performance within your computing environment. With active partition mobility, you can manage workloads with minimal downtime.

The PowerVM Server Evacuation function allows to perform a server evacuation operation. This process is used to move all migration capable logical partitions from one system to another at a time. Any upgrade or maintenance operations can be performed after all the partitions are migrated and the source system is powered off.

You can migrate all the migration capable AIX, and Linux partitions from the source server to the destination server by running the following command from the HMC command line:

```
migrlpar -o m -m source_server -t target_server --all
```

Hardware and operating system requirements for Live Partition Mobility

PowerVM Live Partition Mobility requires a license for PowerVM Enterprise Edition, and it is supported in compliance with all Operating Systems compatible with POWER8 technology.

The Virtual I/O Server partition itself cannot be migrated.

For more information about Live Partition Mobility and how to implement it, see: *IBM PowerVM Live Partition Mobility*, SG24-7460.

3.4.6 Active Memory Sharing

Active Memory Sharing is an IBM PowerVM advanced memory virtualization technology that provides system memory virtualization capabilities to IBM Power Systems, allowing multiple partitions to share a common pool of physical memory.

Active Memory Sharing is available only with the Enterprise version of PowerVM.

The physical memory of an IBM Power System can be assigned to multiple partitions in either dedicated or shared mode. The system administrator has the capability to assign some physical memory to a partition and some physical memory to a pool that is shared by other partitions. A single partition can have either dedicated or shared memory:

- With a pure dedicated memory model, the system administrator's task is to optimize available memory distribution among partitions. When a partition suffers degradation because of memory constraints and other partitions have unused memory, the administrator can manually issue a dynamic memory reconfiguration.
- With a shared memory model, the system automatically decides the optimal distribution of the physical memory to partitions and adjusts the memory assignment based on partition load. The administrator reserves physical memory for the shared memory pool, assigns partitions to the pool, and provides access limits to the pool.

Active Memory Sharing can be used to increase memory utilization on the system either by decreasing the global memory requirement or by allowing the creation of additional partitions on an existing system. Active Memory Sharing can be used in parallel with Active Memory Expansion on a system running a mixed workload of several operating system. For example, AIX partitions can take advantage of Active Memory Expansion. Other operating systems take advantage of Active Memory Sharing also.

For additional information regarding Active Memory Sharing, see *PowerVM Virtualization Active Memory Sharing*, REDP-4470.

3.4.7 Active Memory Deduplication

In a virtualized environment, the systems might have a considerable amount of duplicated information that is stored on RAM after each partition has its own operating system, and some of them might even share the same kinds of applications. On heavily loaded systems, this behavior might lead to a shortage of the available memory resources, forcing paging by the Active Memory Sharing partition operating systems, the Active Memory Deduplication pool, or both, which might decrease overall system performance

Active Memory Deduplication allows the POWER Hypervisor to dynamically map identical partition memory pages to a single physical memory page within a shared memory pool. This way enables a better utilization of the Active Memory Sharing shared memory pool, increasing the system's overall performance by avoiding paging. Deduplication can cause the hardware to incur fewer cache misses, which also leads to improved performance.

Active Memory Deduplication depends on the Active Memory Sharing feature to be available, and consumes CPU cycles donated by the Active Memory Sharing pool's Virtual I/O Server (VIOS) partitions to identify deduplicated pages. The operating systems that are running on the Active Memory Sharing partitions can "hint" to the POWER Hypervisor that some pages (such as frequently referenced read-only code pages) are particularly good for deduplication.

To perform deduplication, the hypervisor cannot compare every memory page in the Active Memory Sharing pool with every other page. Instead, it computes a small signature for each page that it visits and stores the signatures in an internal table. Each time that a page is inspected, a look-up of its signature is done in the known signatures in the table. If a match is found, the memory pages are compared to be sure that the pages are really duplicates. When a duplicate is found, the hypervisor remaps the partition memory to the existing memory page and returns the duplicate page to the Active Memory Sharing pool.

From the LPAR perspective, the Active Memory Deduplication feature is completely transparent. If an LPAR attempts to modify a deduplicated page, the hypervisor grabs a free page from the Active Memory Sharing pool, copies the duplicate page contents into the new page, and maps the LPAR's reference to the new page so that the LPAR can modify its own unique page.

For additional information regarding Active Memory Dedupliation, see *Power Systems Memory Deduplication*, REDP-4827.

3.4.8 Operating system support for PowerVM

At the time of writing all PowerVM features are supported by the operating systems compatible with the POWER8 servers, except the Active Memory Expansion which is not supported with Linux.

Table 3-4 summarizes the PowerVM features that are supported by the operating systems compatible with the POWER8 processor-based servers.

Table 3-4 Virtualization features supported by AIX, and Linux

Feature	AIX 6.1 TL8	AIX 7.1	RHEL 6.5	SLES 11 SP3
Virtual SCSI	Yes	Yes	Yes	Yes
Virtual Ethernet	Yes	Yes	Yes	Yes
Shared Ethernet Adapter	Yes	Yes	Yes	Yes
Virtual Fibre Channel	Yes	Yes	Yes	Yes
Virtual Tape	Yes	Yes	Yes	Yes
Logical partitioning	Yes	Yes	Yes	Yes
DLPAR I/O adapter add/remove	Yes	Yes	Yes	Yes
DLPAR processor add/remove	Yes	Yes	Yes	Yes
DLPAR memory add	Yes	Yes	Yes	Yes

Feature	AIX 6.1 TL8	AIX 7.1	RHEL 6.5	SLES 11 SP3
DLPAR memory remove	Yes	Yes	Yes	Yes
Micro-Partitioning	Yes	Yes	Yes	Yes
Shared dedicated capacity	Yes	Yes	Yes	Yes
Multiple Shared Processor Pools	Yes	Yes	Yes	Yes
Virtual I/O Server	Yes	Yes	Yes	Yes
Integrated Virtualization Manager	Yes	Yes	Yes	Yes
Suspend/resume and Hibernation[a]	Yes	Yes	Yes	Yes
Shared Storage Pools	Yes	Yes	Yes	Yes
Thin provisioning	Yes	Yes	Yes	Yes
Active Memory Sharing[b]	Yes	Yes	Yes	Yes
Active Memory Deduplication	Yes	Yes	Yes	Yes
Live Partition Mobility	Yes	Yes	Yes	Yes
Simultaneous multithreading (SMT)	Yes	Yes	Yes	Yes
Active Memory Expansion	Yes	Yes	No	No

a. Note: At the time of writing, Suspend/Resume is not available. Check with your IBM representative for availability on POWER8 platforms.
b. Note: At the time of writing Active Memory Sharing when used with Live Partition Mobility is not supported. Check with your IBM representative for availability on POWER8 platforms.

For more information about specific features for Linux, see the following web page:

http://pic.dhe.ibm.com/infocenter/lnxinfo/v3r0m0/index.jsp?topic=%2Fliaam%2Fsupportedfeaturesforlinuxonpowersystemsservers.htm

3.4.9 Linux support

The IBM Linux Technology Center (LTC) contributes to the development of Linux by providing support for IBM hardware in Linux distributions. In particular, the LTC has available tools and code to the Linux communities to take advantage of the POWER8 technology and develop POWER8 optimized software.

For more information about specific Linux distributions, see the following web page:

http://pic.dhe.ibm.com/infocenter/lnxinfo/v3r0m0/index.jsp?topic=%2Fliaam%2Fliaamdistros.htm

3.4.10 PowerVM Simplification

The Hardware Management Console code for POWER8 servers must run minimum Version 8 Release 8 (HMC V8R8.1.0) version

Table 3-5 lists the minimum requirement for HMC V8R8.1.0.

Table 3-5 Minimum requirements for HMC V8R8.1.0

Function	Minimum requirement
HMC hardware	7042-C08 (deskside), 7042-CR5 (rack-mount) or later
HMC memory	2 GB
VIOS level	2.2.3.3

- PowerVM simplification improves, and simplify the overall end-to-end management, deployment of virtualization on Power Systems
- Comprehensive effort to simplify PowerVM and VIOS management.
- Simplification for HMC touch point users as well as programmatic consumers
- Simplified and automated initial set up via deploy templates
- No touch VIOS management - full GUI management interface
- Single point of management for all PowerVM configuration
- Integrated performance & capacity metrics

Figure 3-11 on page 99 illustrates new options in the HMC V8R1 available for PowerVM simplification. Manage PowerVM opens a new window, where an user can view and manage all the aspects of a PowerVM configuration such as Shared Ethernet Adapter, Virtual Networks, Virtual Storage using the graphical interface only.

Templates allows specify the configuration details for the system I/O, memory, storage, network, processor, and other partition resources. An user can use the predefined or captured templates that are available in the template library to deploy a new system. Two type of templates are available in the template library:

- System Template contains configuration details for system resources, such as system I/O, memory, storage, processors, network, and physical I/O adapters. You can use these system templates to deploy the system.
- Partition Template contains configuration details for the partition resources. An user can configure Logical Partitions with the predefined templates or by creating a custom templates.

Using these options, an user can deploy a system, select a system template, and click Deploy. After deploying a system, an user can choose a partition template from the library.

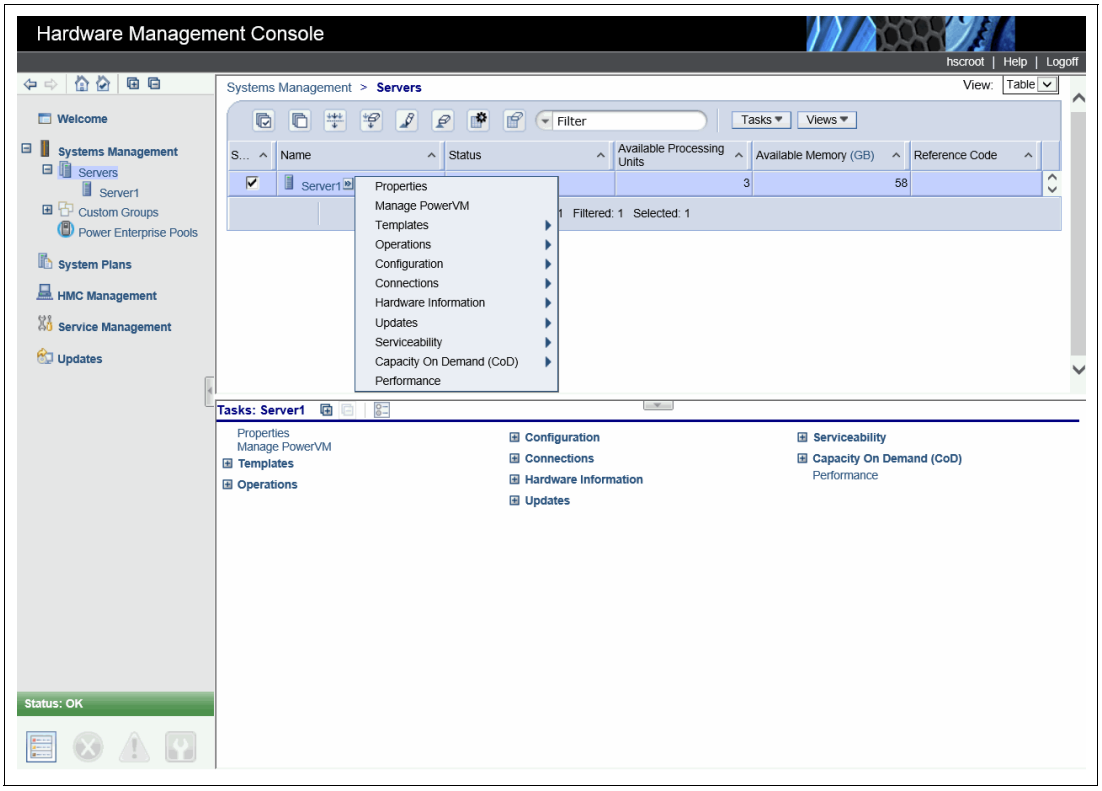


Figure 3-11 PowerVM new level tasks

Performance function opens the Performance and Capacity Monitoring window as shown at the top of the screen in Figure 3-12.

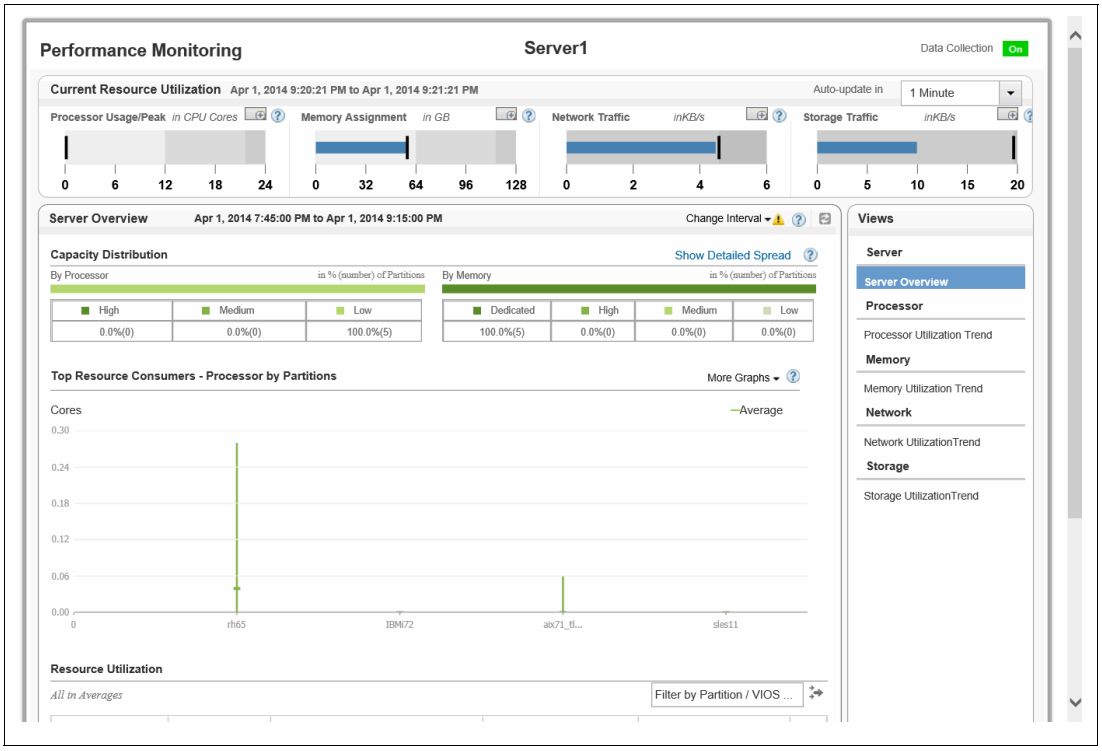


Figure 3-12 HMC performance monitoring CPU, Memory assignment (top of the screen)

Figure 3-13, shows the bottom of the screen where performance and capacity data are presented in a graphical format.

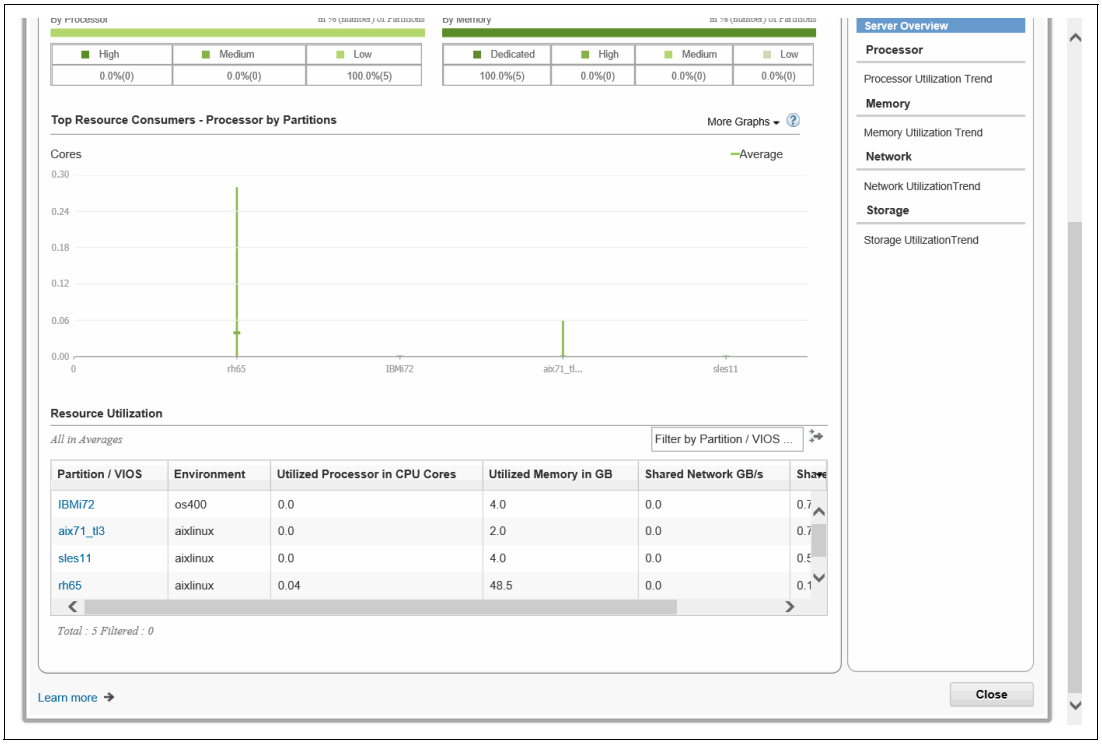


Figure 3-13 HMC performance monitoring CPU, Memory assignment (bottom of the screen)

3.5 System Planning Tool

The IBM System Planning Tool (SPT) helps you design systems to be partitioned with logical partitions. You can also plan for and design non-partitioned systems by using the SPT. The resulting output of your design is called a *system plan*, which is stored in a `.sysplan` file. This file can contain plans for a single system or multiple systems. The `.sysplan` file can be used for the following reasons:

- To create reports
- As input to the IBM configuration tool (e-Config)
- To create and deploy partitions on your system (or systems) automatically

System plans that are generated by the SPT can be deployed on the system by the Hardware Management Console (HMC), or Integrated Virtualization Manager (IVM).

Automatically deploy: Ask your IBM representative or IBM Business Partner to use the Customer Specified Placement manufacturing option if you want to automatically deploy your partitioning environment on a new machine. SPT looks for the resource's allocation to be the same as that specified in your `.sysplan` file.

You can create an entirely new system configuration, or you can create a system configuration based on any of the following items:

- Performance data from an existing system that the new system is to replace
- Performance estimates that anticipates future workloads that you must support
- Sample systems that you can customize to fit your needs

Integration between the System Planning Tool and both the Workload Estimator and IBM Performance Management allows you to create a system that is based on performance and capacity data from an existing system or that is based on new workloads that you specify.

You can use the SPT before you order a system to determine what you must order to support your workload. You can also use the SPT to determine how you can partition a system that you already have.

Using the SPT is an effective way of documenting and backing up key system settings and partition definitions. With it, the user can create records of systems and export them to their personal workstation or backup system of choice. These same backups can then be imported back onto the same managed console when needed. This step can be useful when cloning systems. enabling the user to import the system plan to any managed console multiple times.

The SPT and its supporting documentation is on the IBM System Planning Tool site:

http://www.ibm.com/systems/support/tools/systemplanningtool/

3.6 IBM PowerVC

IBM Power Virtualization Center (IBM PowerVC) is designed to simplify the management of virtual resources in your Power Systems environment.

Once the product code is loaded, IBM PowerVC's no-menus interface will guide you through three simple configuration steps to register physical hosts, storage providers, and network resources and start capturing and intelligently deploying your VMs among other tasks shown in the following list:

- Create virtual machines and then resize and attach volumes to them.
- Import existing virtual machines and volumes so they can be managed by IBM PowerVC.
- Monitor the utilization of the resources that are in your environment.
- Migrate virtual machines while they are running (hot migration).
- Deploy images quickly to create new virtual machines that meet the demands of your ever-changing business needs.

IBM PowerVC is built on OpenStack. OpenStack is an open source software that controls large pools of server, storage, and networking resources throughout a datacenter.

IBM PowerVC is available in two editions:

- IBM Power Virtualization Center Express Edition
- IBM Power Virtualization Center Standard Edition

Table 3-6 shows an overview of the key features included with IBM PowerVC Editions:

Table 3-6 IBM PowerVC Editions key features overview.

IBM Power Virtualization Center Express Edition	IBM Power Virtualization Center Standard Edition
► Supports IBM Power Systems hosts that are managed by the Integrated Virtualization Manager (IVM). ► Supports storage area networks, local storage, and a combination in the same environment. ► Supports a single Virtual I/O Server virtual machine on each host.	► Supports IBM Power Systems hosts that are managed by a Hardware Management Console (HMC). ► Supports storage area networks. ► Supports multiple Virtual I/O Servers virtual machines on each host.

For more information on IBM PowerVC, refer to the IBM PowerVC Introduction and Configuration, SG24-8199, available at:

http://www.redbooks.ibm.com/redpieces/abstracts/sg248199.html?Open

3.7 IBM PowerVP

IBM Power Virtualization Performance (PowerVP) for Power Systems is a new product that offers a performance view into an IBM PowerVM virtualized environment running on the latest firmware of IBM Power systems. It is capable of showing which virtual workloads are using specific physical resources on an IBM Power Systems server.

IBM PowerVP helps reduce time and complexity to find and display performance bottlenecks through a simple dashboard that shows the performance health of the system. It can help simplify both prevention and troubleshooting and thus reduce the cost of performance management.

It assists you in the following way:

- Shows workloads in real-time highlighting of possible problems or bottlenecks (overcommitted resources)
- Helps better utilize virtualized IBM Power System servers by showing distribution of workload
- Is able to replay saved historical data
- Helps with the resolution of performance related issues
- Helps to proactively address future issues that could affect performance

IBM PowerVP is integrated with the POWER Hypervisor and collects performance data directly from PowerVM Hypervisor, which offers the most accurate performance information about virtual machines running on IBM Power Systems. This performance information is displayed on a real-time, continuous GUI dashboard and it is also available for historical review.

Features of IBM PowerVP include:

- Real-time, continuous graphical monitor (dashboard) that delivers an easy-to-read display showing the overall performance health of the Power server.
- Customizable performance thresholds that enables you to customize the dashboard to match your monitoring requirements.
- Historical statistics that enables you to go back in time and replay performance data sequences to find out performance bottlenecks.
- System-level performance views that show all LPARs (Virtual Machines) and how they are using real system resources.
- Virtual machine drill-down, which gives you more performance details for each virtual machine, displaying detailed information on various resources like CPU, memory, and disk activity.
- Support for all virtual machine types including AIX, IBM i, and Linux.
- Background data collection, which enables performance data to be collected when the GUI is not active.

IBM PowerVP even allows an administrator to drill down and view specific adapter, bus or CPU usage. An administrator can see the hardware adapters and how much workload is placed on them. IBM PowerVP provides both an overall and detailed view of IBM Power System server hardware so it is easy to see how virtual machines are consuming resources. More information about PowerVP can be found at:

http://www.ibm.com/systems/power/software/performance/

3.8 VIOS 2.2.3 features

Shared Storage Pools have been enhanced to improve flexibility, scalability, and resiliency. These improvements include mirroring of the storage pool, dynamic contraction of the storage pool, dynamic disk growth within the storage pool, and scaling improvements. Virtual I/O Server Performance Advisor has been enhanced to provide support for NPIV and Fibre Channel, Virtual Networking and Shared Ethernet Adapter, and Shared Storage Pool configurations. Shared Ethernet Adapter Failover configuration has been simplified by removing the extra complexity when configuring a control channel. The functionality of the Virtual I/O Server has been improved along with Live Partition Mobility performance

- Simplified Shared Ethernet Adapter Failover configuration setup
- Shared Storage Pools enhancements
- Shared Ethernet Adapter by default uses largesend attribute
- The Virtual I/O Server Performance Advisor has been enhanced to provide support for NPIV and Fibre Channel, Virtual Networking, Shared Ethernet Adapter, and Shared Storage Pool configurations.
- Support for Live Partition Mobility performance enhancements to better utilize 10 Gb Ethernet Adapters used within mobility process, and PowerVM server evacuation function.

4

Continuous availability and manageability

This chapter provides information about IBM reliability, availability, and serviceability (RAS) design and features. This set of technologies, implemented on IBM Power Systems servers, improves your architecture's total cost of ownership (TCO) by reducing planned and unplanned down time.

The elements of RAS can be described as follows:

- Reliability: Indicates how infrequently a defect or fault in a server occurs
- Availability: Indicates how infrequently the functionality of a system or application is impacted by a fault or defect
- Serviceability: Indicates how well faults and their effects are communicated to system managers and how efficiently and non disruptively the faults are repaired

Each successive generation of IBM servers is designed to be more reliable than the previous server family. POWER8 processor-based servers have new features to support new levels of virtualization, help ease administrative burden, and increase system utilization.

Reliability starts with components, devices, and subsystems designed to be fault-tolerant. POWER8 uses lower voltage technology, improving reliability with stacked latches to reduce soft error susceptibility. During the design and development process, subsystems go through rigorous verification and integration testing processes. During system manufacturing, systems go through a thorough testing process to help ensure high product quality levels.

The processor and memory subsystem contain features that are designed to avoid or correct environmentally induced, single-bit, intermittent failures. The features can also handle solid faults in components, including selective redundancy to tolerate certain faults without requiring an outage or parts replacement.

© Copyright IBM Corp. 2014. All rights reserved.

4.1 Reliability

Highly reliable systems are built with highly reliable components. On IBM POWER processor-based systems, this basic principle is expanded upon with a clear design for reliability architecture and methodology. A concentrated, systematic, architecture-based approach is designed to improve overall system reliability with each successive generation of system offerings.

4.1.1 Designed for reliability

Systems that are designed with fewer components and interconnects have fewer opportunities to fail. Simple design choices such as integrating processor cores on a single POWER chip can dramatically reduce the opportunity for system failures. The POWER8 chip has more cores per Processor module and also the I/O Hub Controller is integrated in the Processor core. That generates a PCIe BUS directly from the Processor module.

Parts selection also plays a critical role in overall system reliability. IBM uses three grades of components with grade 3 being defined as industry standard (“off-the-shelf” components). As shown in Figure 4-1, using stringent design criteria and an extensive testing program, the IBM manufacturing team can produce grade 1 components that are expected to be 10 times more reliable than industry standard. Engineers select grade 1 parts for the most critical system components. Newly introduced organic packaging technologies, rated grade 5, achieve the same reliability as grade 1 parts.

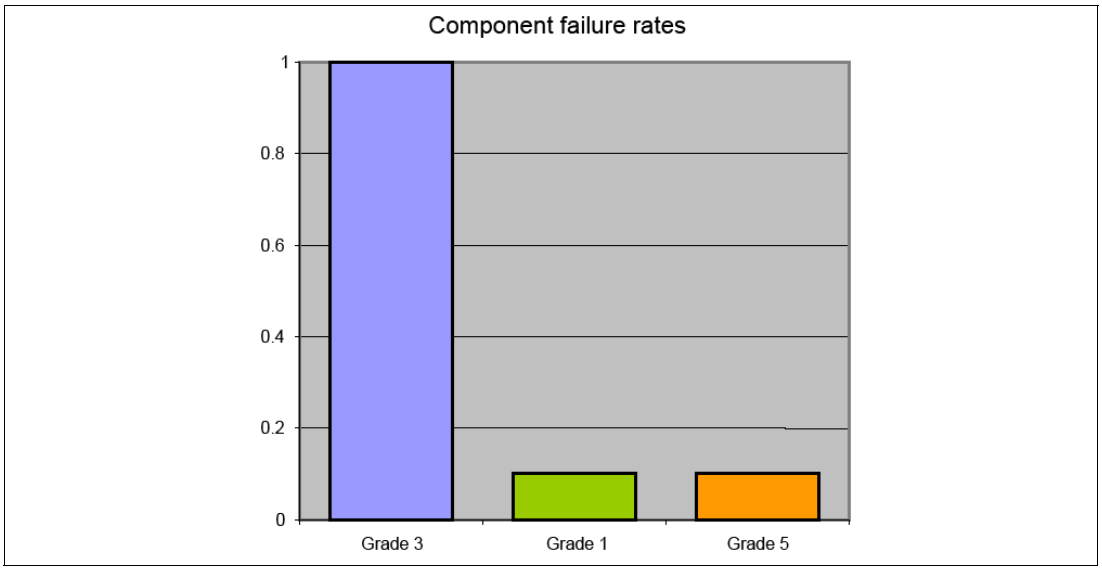


Figure 4-1 Component failure rates

4.1.2 Placement of components

Packaging is designed to deliver both high performance and high reliability. For example, the reliability of electronic components is directly related to their thermal environment. That is, large decreases in component reliability are directly correlated with relatively small increases in temperature. All POWER processor-based systems are carefully packaged to ensure adequate cooling. Critical system components such as the POWER8 processor chips are positioned on the planar so that they receive clear air flow during operation. In addition, POWER processor-based systems are built with redundant, variable-speed fans that can

automatically increase output to compensate for increased heat in the central electronic complex.

4.1.3 Redundant components and concurrent repair

High-opportunity components (those that most affect system availability) are protected with redundancy and the ability to be repaired concurrently.

The use of these redundant components allows the system to remain operational:

- POWER8 cores, which include redundant bits in L1 instruction and data caches, L2 caches, and L2 and L3 directories
- POWER8 Processors Memory Buffer includes also a L4 Cache capability with similar error protection capabilities as available in the L3 Cache
- Power S822 main memory DIMMs, which use an innovative ECC algorithm, from IBM research, that improves bit-error correction and memory failures
- Redundant and hot-swap cooling
- Redundant and hot-swap power supplies

For maximum availability, be sure to connect power cords from the same system to two separate power distribution units (PDUs) in the rack, and to connect each PDU to independent power sources. Tower form factor power cords must be plugged into two independent power sources to achieve maximum availability.

Before ordering: Check your configuration for optional redundant components before ordering your system.

4.2 Availability

First-failure data capture (FFDC) is the capability of IBM hardware and microcode to continuously monitor hardware functions. This process includes predictive failure analysis, which is the ability to track intermittent correctable errors and to take components offline before they reach the point of hard failure. This way avoids causing a system outage.

The POWER8 family of systems can do the following automatic functions:

- Self-diagnose and self-correct errors during run time.
- Automatically reconfigure to mitigate potential problems from suspect hardware.
- Self-heal or automatically substitute good components for failing components.

Remember: Error detection and fault isolation is independent of the operating system in POWER8 processor-based servers.

This chapter describes IBM POWER8 processor-based systems technologies, focused on keeping a system running. For a specific set of functions that are focused on detecting errors before they become serious enough to stop computing work, see 4.3.1, “Detecting” on page 116.

4.2.1 Partition availability priority

POWER8 systems can assign availability priorities to partitions. If the system detects that a processor core is about to fail, it is taken offline. If the partitions on the system require more processor units than remain in the system, the firmware determines which partition has the lowest priority and attempts to claim the needed resource. On a properly configured POWER processor-based server, this capability allows the system manager to ensure that capacity is first obtained from a low-priority partition instead of a high-priority partition.

This capability gives the system an additional stage before an unplanned outage. If insufficient resources exist to maintain full system availability, the server attempts to maintain partition availability according to user-defined priority.

Partition availability priority is assigned to partitions by using a *weight value* or integer rating. The lowest priority partition is rated at 0 (zero) and the highest priority partition is rated at 255. The default value is set to 127 for standard partitions and 192 for Virtual I/O Server (VIOS) partitions. You can vary the priority of individual partitions through the hardware management console.

4.2.2 General detection and deallocation of failing components

Runtime correctable or recoverable errors are monitored to determine whether there is a pattern of errors. If these components reach a predefined error limit, the service processor initiates an action to deconfigure the faulty hardware, helping to avoid a potential system outage and to enhance system availability.

Persistent deallocation

To enhance system availability, a component that is identified for deallocation or deconfiguration on a POWER processor-based system is flagged for persistent deallocation. Component removal can occur either dynamically (while the system is running) or at boot time (IPL), depending both on the type of fault and when the fault is detected.

In addition, unrecoverable hardware faults can be deconfigured from the system after the first occurrence. The system can be rebooted immediately after failure and resume operation on the remaining stable hardware. This way prevents the faulty hardware from affecting system operation again; the repair action is deferred to a more convenient, less critical time.

The following components have the capability to be persistently deallocated:

- Processor
- L2 and L3 cache lines (Cache lines are dynamically deleted.)
- Memory
- Deconfigure or bypass failing I/O adapters

Processor instruction retry

As introduced with the POWER6 technology, the POWER8 processor can retry processor instructions and do alternate processor recovery for several core-related faults. In this way, exposure to both permanent and intermittent errors in the processor core are significantly reduced.

Intermittent errors are generally not repeatable, often because of cosmic rays or other sources of radiation.

With the instruction retry function, when an error is encountered in the core, in caches and certain logic functions, the POWER8 processor first automatically retries the instruction. If the source of the error was truly transient, the instruction succeeds and the system can continue as before.

Before POWER6: On IBM systems prior to POWER6, such an error typically caused a checkstop

Alternate processor retry

Hard failures are more difficult; they are permanent errors that are replicated each time that the instruction is repeated. Retrying the instruction does not help in this situation because the instruction will continue to fail.

As introduced with POWER6, POWER8 processors can extract the failing instruction from the faulty core and retry it elsewhere in the system. The failing core is then dynamically deconfigured and scheduled for replacement.

Dynamic processor deallocation

Dynamic processor deallocation enables automatic deconfiguration of processor cores when patterns of recoverable core-related faults are detected. Dynamic processor deallocation prevents a recoverable error from escalating to an unrecoverable system error, which might otherwise result in an unscheduled server outage. Dynamic processor deallocation relies on the service processor's ability to use FFDC-generated recoverable error information to notify the POWER Hypervisor when a processor core reaches its predefined error limit. The POWER Hypervisor then dynamically deconfigures the failing core and notifies the system administrator that a replacement is needed. The entire process is transparent to the partition owning the failing instruction.

Single processor checkstop

As in the POWER6 processor, the POWER8 processor provides single core check-stopping for certain processor logic, command, or control errors that cannot be handled by the availability enhancements in the preceding section.

This approach significantly reduces the probability of any one processor affecting total system availability by containing most processor checkstops to the partition that was using the processor at the time that full checkstop goes into effect.

Even with all these availability enhancements to prevent processor errors from affecting system-wide availability, errors might occur that can result in a system-wide outage.

4.2.3 Memory protection

A memory protection architecture that provides good error resilience for a relatively small L1 cache might be inadequate for protecting the much larger system main store. Therefore, a variety of protection methods are used in all POWER processor-based systems to avoid uncorrectable errors in memory.

Memory protection plans must account for many factors, including the following factors:

- Size
- Desired performance
- Memory array manufacturing characteristics

POWER8 processor-based systems have various protection schemes designed to prevent, protect, or limit the effect of errors in main memory:

- Chipkill

  Chipkill is an enhancement that enables a system to sustain the failure of an entire DRAM chip. An ECC word uses 18 DRAM chips from two DIMM pairs, and a failure on any of the DRAM chips can be fully recovered by the ECC algorithm. The system can continue indefinitely in this state with no performance degradation until the failed DIMM can be replaced.

- 72-byte ECC

  In POWER8, an ECC word consists of 72 bytes of data. Of these, 64 bytes are used to hold application data. The remaining eight bytes are used to hold check bits and additional information about the ECC word.DIMMs designed by IBM with a memory buffer on each of the DIMM and DRAM modules for holding data, doing error checking and correcting plus spare DRAM modules to allow a failed DRAM module to be replaced with a spare to avoid replacing a DIMM for such a failure. This a a improvement over POWER7 based one socket and two socket servers which had the same level of ECC but no spare DRAMs.

- Hardware scrubbing

  Hardware scrubbing is a method used to handle intermittent errors. IBM POWER processor-based systems periodically address all memory locations. Any memory locations with a correctable error are rewritten with the correct data.

- Cyclic redundancy check (CRC)

  The bus that is transferring data between the processor and the memory uses CRC error detection with a failed operation-retry mechanism and the ability to dynamically retune the bus parameters when a fault occurs. In addition, the memory bus has spare capacity to substitute a data bit-line whenever it is determined to be faulty.

- Memory Channel Repair

  The memory channel design includes a CRC error checking capability. This includes the ability to re-try a failed bus operation and to re-train the channel when excessive CRC errors are seen.

  The design includes the ability to dynamically replace one of the bits on the bus (dynamic bit-lane sparing) based on a hardware detected error. The firmware and hardware do not support detection of which bit is at fault when there is a CRC error other than as part of a re-train operation. Therefore, the POWER8 System does not support dynamic bit-lane sparing based on firmware detecting too many re-training (or channel init) operations.

POWER8 memory subsystem

The POWER8 processor chip contains two memory controllers with four DMI channels per memory controller. Each channel connects to a single DIMM, A processor chip can address eight CDIMM modules.

The bus transferring data between the processor and the memory uses CRC error detection with a failed operation retry mechanism and the ability to dynamically retune bus parameters when a fault occurs. In addition, the memory bus has spare capacity to substitute a spare data bit-line for one that is determined to be faulty.

Advanced memory buffer chips are exclusive to IBM and help to increase performance, acting as read/write buffers. The memory buffer contains a L4 cache capability with error protection capabilities similar to those of a processor L3 cache.

Memory page deallocation

Although coincident cell errors in separate memory chips are statistically rare, IBM POWER8 processor-based systems can contain these errors by using a memory page deallocation scheme for partitions that are running IBM AIX operating system, and also for memory pages that are owned by the POWER Hypervisor. If a memory address experiences an uncorrectable or repeated correctable single cell error, the service processor sends the memory page address to the POWER Hypervisor to be marked for deallocation.

Pages that are used by the POWER Hypervisor are deallocated as soon as the page is released. In other cases, the POWER Hypervisor notifies the owning partition that the page must be deallocated. Where possible, the operating system moves any data currently contained in that memory area to another memory area and removes the pages associated with this error from its memory map, no longer addressing these pages. The operating system performs memory page deallocation without any user intervention and is transparent to users and applications.

The POWER Hypervisor maintains a list of pages marked for deallocation during the current platform initial program load (IPL). During a partition IPL, the partition receives a list of all the bad pages in its address space. In addition, if memory is dynamically added to a partition (through a dynamic LPAR operation), the POWER Hypervisor warns the operating system when memory pages are included that need to be deallocated.

Finally, if an uncorrectable error in memory is discovered, the logical memory block that is associated with the address that has the uncorrectable error is marked for deallocation by the POWER Hypervisor. This deallocation becomes effective on a partition reboot if the logical memory block is assigned to an active partition at the time of the fault.

In addition, the system will deallocate the entire memory group that is associated with the error on all subsequent system reboots until the memory is repaired. This precaution is intended to guard against future uncorrectable errors while waiting for parts replacement.

Memory persistent deallocation

Defective memory that is discovered at boot time is automatically switched off. If the service processor detects a memory fault at boot time, it marks the affected memory as bad so that it is not used on subsequent reboots.

Upon reboot, if not enough memory is available to meet minimum partition requirements, the POWER Hypervisor will reduce the capacity of one or more partitions.

Depending on the configuration of the system, the HMC IBM Service Focal Point™, OS Service Focal Point, or service processor will receive a notification of the failed component, and will trigger a service call.

4.2.4 Cache protection

POWER8 processor-based systems are designed with cache protection mechanisms, including cache-line delete in both L2 and L3 arrays, processor instruction retry and alternate processor recovery protection on L1-I and L1-D, and redundant “repair” bits in L1-I, L1-D, and L2 caches, and L2 and L3 directories.

L1 instruction and data array protection

The POWER8 processor instruction and data caches are protected against intermittent errors using processor instruction retry and against permanent errors by alternate processor recovery, both mentioned previously. L1 cache is divided into sets. POWER8 processor can deallocate all but one before doing a processor instruction retry.

In addition, faults in the Segment Lookaside Buffer (SLB) array are recoverable by the POWER Hypervisor. The SLB is used in the core to do address translation calculations.

L2 and L3 array protection

The L2 and L3 caches in the POWER8 processor are protected with double-bit detect single-bit correct error detection code (ECC). Single-bit errors are corrected before forwarding to the processor and are subsequently written back to the L2 and L3 cache.

In addition, the caches maintain a cache-line delete capability. A threshold of correctable errors that are detected on a cache line can result in the data in the cache line being purged and the cache line removed from further operation without requiring a reboot. An ECC uncorrectable error detected in the cache can also trigger a purge and delete of the cache line. This results in no loss of operation because an unmodified copy of the data can be held on system memory to reload the cache line from main memory. Modified data is handled through Special Uncorrectable Error handling.

L2 and L3 deleted cache lines are marked for persistent deconfiguration on subsequent system reboots until they can be replaced.

L4 cache protection

The POWER8 processor has an integrated memory buffer with L4 cache error protection similar to a L3 cache error protection.

4.2.5 Special Uncorrectable Error handling

Although rare, an uncorrectable data error can occur in memory or cache. IBM POWER processor-based systems attempt to limit the impact of an uncorrectable error to the least possible disruption, using a well-defined strategy that first considers the data source. Sometimes, an uncorrectable error is temporary in nature and occurs in data that can be recovered from another repository, as in the following example:

- Data in the instruction L1 cache is never modified within the cache itself. Therefore, an uncorrectable error discovered in the cache is treated like an ordinary cache miss, and correct data is loaded from the L2 cache.

- The L2 and L3 cache of the POWER8 processor-based systems can hold an unmodified copy of data in a portion of main memory. In this case, an uncorrectable error simply triggers a reload of a cache line from main memory.

In cases where the data cannot be recovered from another source, a technique named Special Uncorrectable Error (SUE) handling is used to prevent an uncorrectable error in memory or cache from immediately causing the system to terminate. That is, the system tags the data and determines whether it will ever be used again:

- If the error is irrelevant, SUE will not force a checkstop.
- If data is used, termination can be limited to the program/kernel or hypervisor that owns the data, or freeze the I/O adapters that are controlled by an I/O hub controller if data is going to be transferred to an I/O device.

When an uncorrectable error is detected, the system modifies the associated ECC word, thereby signaling to the rest of the system that the “standard” ECC is no longer valid. The service processor is then notified and takes appropriate actions. When running AIX 5.2, or later, or Linux, and a process attempts to use the data, the operating system is informed of the error and might terminate, or only terminate a specific process that is associated with the corrupt data, depending on the operating system and firmware level and whether the data was associated with a kernel or non-kernel process.

Only in the case where the corrupt data is used by the POWER Hypervisor must the entire system be rebooted, thereby preserving overall system integrity.

Depending on system configuration and the source of the data, errors encountered during I/O operations might not result in a machine check. Instead, the incorrect data is handled by the processor host bridge (PHB) chip. When the PHB chip detects a problem, it rejects the data, preventing data from being written to the I/O device.

The PHB then enters a freeze mode, halting normal operations. Depending on the model and type of I/O being used, the freeze might include the entire PHB chip, or simply a single bridge, resulting in the loss of all I/O operations that use the frozen hardware until a power-on reset of the PHB is done. The impact to partitions depends on how the I/O is configured for redundancy. In a server configured for failover availability, redundant adapters spanning multiple PHB chips can enable the system to recover transparently, without partition loss.

4.2.6 PCI Enhanced Error Handling

IBM estimates that PCI adapters can account for a significant portion of the hardware-based errors on a large server. Although servers that rely on boot-time diagnostics can identify failing components to be replaced by hot-swap and reconfiguration, runtime errors pose a more significant problem.

PCI adapters are generally complex designs involving extensive on-board instruction processing, often on embedded microcontrollers. They tend to use industry standard grade components with an emphasis on product cost relative to high reliability. In certain cases, they might be more likely to encounter internal microcode errors or many of the hardware errors described for the rest of the server.

The traditional means of handling these problems is through adapter internal error reporting and recovery techniques in combination with operating system device driver management and diagnostics. In certain cases, an error in the adapter might cause transmission of bad data on the PCI bus itself, resulting in a hardware-detected parity error and causing a global machine-check interrupt, eventually requiring a system reboot to continue.

PCI Enhanced Error Handling (EEH) enabled adapters respond to a special data packet that is generated from the affected PCI slot hardware by calling system firmware, which will examine the affected bus, allow the device driver to reset it, and continue without a system reboot. For Linux, EEH support extends to the majority of frequently used devices, although various third-party PCI devices might not provide native EEH support.

Each processor module can directly drive two I/O slots or devices with the PCIe controllers in each processor and with an external module of any kind, or a single controller can handle additional functions.

While the I/O Hub has been integrated in the POWER8 processor module it still retains a design that supports end-point error recovery as well as a 'freeze on fault' behavior and fault isolation so errors can be contained to a partition using the I/O.

4.3 Serviceability

IBM Power Systems design considers both IBM and the client's needs. The IBM Serviceability Team, enhanced the base service capabilities and continues to implement a strategy that incorporates best-of-its-kind service characteristics from diverse IBM Systems offerings.

The purpose of serviceability is to repair the system while attempting to minimize or eliminate service cost (within budget objectives), while maintaining high customer satisfaction. Serviceability includes system installation, MES (system upgrades/downgrades), and system maintenance/repair. Depending on the system and warranty contract, service may be performed by the customer, an IBM representative, or an authorized warranty service provider.

The serviceability features that are delivered in this system provide a highly efficient service environment by incorporating the following attributes:

- Design for customer setup (CSU), customer installed features (CIF), and customer-replaceable units (CRU)
- Error detection and fault isolation (ED/FI)
- First-failure data capture (FFDC)
- Converged service approach across multiple IBM server platforms

By delivering on these goals, IBM Power Systems servers enable faster and more accurate repair, and reduce the possibility of human error.

Client control of the service environment extends to firmware maintenance on all of the POWER processor-based systems. This strategy contributes to higher systems availability with reduced maintenance costs.

This section provides an overview of the progressive steps of error detection, analysis, reporting, notifying, and repairing found in all POWER processor-based systems.

4.3.1 Detecting

The first and most crucial component of a solid serviceability strategy is the ability to accurately and effectively detect errors when they occur. Although not all errors are a guaranteed threat to system availability, those that go undetected can cause problems because the system has no opportunity to evaluate and act if necessary. Power processor-based systems employ IBM System z® server-inspired error detection

mechanisms, extending from processor cores and memory to power supplies and hard drives.

Service processor

The service processor is a microprocessor that is powered separately from the main instruction processing complex. The service processor provides the capabilities for the following items:

- POWER Hypervisor (system firmware) and Hardware Management Console connection surveillance
- Several remote power control options
- Reset and boot features
- Environmental monitoring

The service processor monitors the server's built-in temperature sensors, sending instructions to the system fans to increase rotational speed when the ambient temperature is above the normal operating range. By using an architected operating system interface, the service processor notifies the operating system of potential environmentally related problems so that the system administrator can take appropriate corrective actions before a critical failure threshold is reached. The service processor can also post a warning and initiate an orderly system shutdown in the following circumstances:

- The operating temperature exceeds the critical level (for example, failure of air conditioning or air circulation around the system).
- The system fan speed is out of operational specification (for example, because of multiple fan failures).
- The server input voltages are out of operational specification.

The service processor can immediately shut down a system in the following circumstances:

- Temperature exceeds the critical level or remains above the warning level for too long.
- Internal component temperatures reach critical levels.
- Non-redundant fan failures occur.

The service processor provides the following features:

- Placing calls

  On systems without a Hardware Management Console, the service processor can place calls to report surveillance failures with the POWER Hypervisor, critical environmental faults, and critical processing faults even when the main processing unit is inoperable.

- Mutual surveillance

  The service processor monitors the operation of the firmware during the boot process, and also monitors the hypervisor for termination. The hypervisor monitors the service processor and can perform a reset and reload if it detects the loss of the service processor. If the reset/reload operation does not correct the problem with the service processor, the hypervisor notifies the operating system; the operating system can then take appropriate action, including calling for service.

- Availability

  The POWER8 family of systems continues to offer and introduce significant enhancements designed to increase system availability.

  As in POWER6, POWER6+, POWER7, POWER7+, and the POWER8 processor has the ability to do processor instruction retry and alternate processor recovery for several core-related faults. This ability significantly reduces exposure to both hard (logic) and soft

(transient) errors in the processor core. Soft failures in the processor core are transient (intermittent) errors, often because of cosmic rays or other sources of radiation, and generally are not repeatable. When an error is encountered in the core, the POWER8 processor first automatically retries the instruction. If the source of the error was truly transient, the instruction succeeds and the system continues as before. On IBM systems before POWER6, this error caused a checkstop.

Hard failures are more difficult; they are true logical errors that are replicated each time the instruction is repeated. Retrying the instruction does not help in this situation. As in POWER6, POWER6+, POWER7, and POWER7+ all POWER8 processors can extract the failing instruction from the faulty core and retry it elsewhere in the system for several faults, after which the failing core is dynamically deconfigured and called out for replacement. These systems are designed to avoid a full system outage.

- Uncorrectable error recovery

  The auto-restart (reboot) option, when enabled, can reboot the system automatically following an unrecoverable firmware error, firmware hang, hardware failure, or environmentally induced (AC power) failure.

The auto-restart (reboot) option must be enabled from the Advanced System Management Interface (ASMI) or from the Control (Operator) Panel.

- Partition availability priority

  Availability priorities can be assigned to partitions. If an alternate processor recovery event requires spare processor resources to protect a workload, when no other means of obtaining the spare resources is available, the system determines which partition has the lowest priority and attempts to claim the needed resource. On a properly configured POWER8 processor-based server, this way allows that capacity to be first obtained from, for example, a test partition instead of a financial accounting system.

- POWER8 cache availability

  The L2 and L3 caches in the POWER8 processor are protected with double-bit detect, single-bit correct error detection code (ECC). In addition, the caches maintain a cache line delete capability. A threshold of correctable errors detected on a cache line can result in the data in the cache line being purged and the cache line removed from further operation without requiring a reboot. An ECC uncorrectable error detected in the cache can also trigger a purge and delete operation of the cache line. This step results in no loss of operation if the cache line contained data that is unmodified from what was stored in system memory. Modified data would be handled through Special Uncorrectable Error handling. L1 data and instruction caches also have a retry capability for intermittent error and a cache set delete mechanism for handling solid failures. In addition, the POWER8 processors also have the ability to dynamically substitute a faulty bit-line in an L3 cache dedicated to a processor with a spare bit-line.For soft errors in caches, key design elements include ECC in the L2 and L3 caches, plus a retry mechanism to handle L1 cache faults.

  For some persistent errors in the processor core, alternate processor recovery allows workload running on one core to be migrated over to another core without taking any applications down in the process. This technique does require co-operation of the PowerVM hypervisor, but with proper virtualization and with sufficient spare capacity can be transparent to operating systems and applications.

  Persistent recoverable errors in the L1/L2 and L3 caches can be handled first by removing from use the portion of the cache containing the error.

  Even when uncorrectable errors occur in caches and may cause some sort of application or other code interruption, cache line-delete can prevent repeat faults without needing to replace any hardware.

- Fault monitoring

  Built-in self-test (BIST) checks processor, cache, memory, and associated hardware that is required for proper booting of the operating system, when the system is powered on at the initial installation or after a hardware configuration change (for example, an upgrade). If a non-critical error is detected or if the error occurs in a resource that can be removed from the system configuration, the booting process is designed to proceed to completion. The errors are logged in the system nonvolatile random access memory (NVRAM). When the operating system completes booting, the information is passed from the NVRAM to the system error log where it is analyzed by error log analysis (ELA) routines. Appropriate actions are taken to report the boot-time error for subsequent service, if required.

- Concurrent access to the service processors menus of the ASMI

  This access allows non disruptive abilities to change system default parameters, interrogate service processor progress and error logs, and set and reset server indicators (Guiding Light for midrange and high-end servers, Light Path for low-end servers), accessing all service processor functions without having to power down the system to the standby state. This allows the administrator or service representative to dynamically access the menus from any web browser-enabled console that is attached to the Ethernet service network, concurrently with normal system operation.

- Managing the interfaces for connecting uninterruptible power source systems to the POWER processor-based systems, performing timed power-on (TPO) sequences, and interfacing with the power and cooling subsystem

Error checkers

IBM POWER processor-based systems contain specialized hardware detection circuitry that is used to detect erroneous hardware operations. Error checking hardware ranges from parity error detection coupled with processor instruction retry and bus retry, to ECC correction on caches and system buses.

All IBM hardware error checkers have distinct attributes:

- Continuous monitoring of system operations to detect potential calculation errors.
- Attempts to isolate physical faults based on runtime detection of each unique failure.
- Ability to initiate a wide variety of recovery mechanisms designed to correct the problem. The POWER processor-based systems include extensive hardware and firmware recovery logic.

Fault isolation registers

Error-checker signals are captured and stored in hardware fault isolation registers (FIRs). The associated logic circuitry is used to limit the domain of an error to the first checker that encounters the error. In this way, runtime error diagnostics can be deterministic so that for every check station, the unique error domain for that checker is defined and documented. Ultimately, the error domain becomes the field-replaceable unit (FRU) call, and manual interpretation of the data is not normally required.

First-failure data capture

First-failure data capture (FFDC) is an error isolation technique. It ensures that when a fault is detected in a system through error checkers or other types of detection methods, the root cause of the fault will be captured without the need to re-create the problem or run an extended tracing or diagnostics program.

For the vast majority of faults, a good FFDC design means that the root cause is detected automatically without intervention by a service representative. Pertinent error data related to the fault is captured and saved for analysis. In hardware, FFDC data is collected from the fault

isolation registers and from the associated logic. In firmware, this data consists of return codes, function calls, and so forth.

FFDC *check stations* are carefully positioned within server logic and data paths to ensure that potential errors can be quickly identified and accurately tracked to a FRU.

This proactive diagnostic strategy is a significant improvement over the classic, less accurate *reboot and diagnose* service approaches.

Fault isolation

The service processor interprets error data that is captured by the FFDC checkers (saved in the FIRs or other firmware-related data capture methods) to determine the root cause of the error event.

Root cause analysis might indicate that the event is recoverable, meaning that a service action point or need for repair has not been reached. Alternatively, it might indicate that a service action point was reached, where the event exceeded a predetermined threshold or was unrecoverable. Based on the isolation analysis, recoverable error-threshold counts can be incremented. No specific service action is necessary when the event is recoverable.

When the event requires a service action, additional required information is collected to service the fault. For unrecoverable errors or for recoverable events that meet or exceed their service threshold (meaning that a service action point was reached) a request for service is initiated through an error logging component.

4.3.2 Diagnosing

General diagnostic objectives are to detect and identify problems so that they can be resolved quickly. IBM diagnostics strategy includes the following elements:

- Provide a common error code format equivalent to a system reference code, system reference number, checkpoint, or firmware error code.
- Provide fault detection and problem isolation procedures. Support remote connection ability to be used by the IBM Remote Support Center or IBM Designated Service.
- Provide interactive intelligence within the diagnostics with detailed online failure information while connected to IBM back-end system.

Using the extensive network of advanced and complementary error detection logic that is built directly into hardware, firmware, and operating systems, the IBM Power Systems servers can perform considerable self-diagnosis.

Because of the FFDC technology that is designed into IBM servers, re-creating diagnostics for failures or requiring user intervention is not necessary. Solid and intermittent errors are designed to be correctly detected and isolated at the time that the failure occurs. Runtime and boot time diagnostics fall into this category.

Boot time

When an IBM Power Systems server powers up, the service processor initializes the system hardware. Boot-time diagnostic testing uses a multitier approach for system validation, starting with managed low-level diagnostics that are supplemented with system firmware initialization and configuration of I/O hardware, followed by OS-initiated software test routines. Boot-time diagnostic routines include the following items:

- Built-in self-tests (BISTs) for both logic components and arrays ensure the internal integrity of components. Because the service processor assists in performing these tests,

the system is enabled to perform fault determination and isolation, whether or not the system processors are operational. Boot-time BISTs can also find faults undetectable by processor-based power-on self-test (POST) or diagnostics.

- Wire-tests discover and precisely identify connection faults between components such as processors, memory
- Initialization of components such as ECC memory, typically by writing patterns of data and allowing the server to store valid ECC data for each location, can help isolate errors.

To minimize boot time, the system determines which of the diagnostics are required to be started to ensure correct operation, based on the way that the system was powered off, or on the boot-time selection menu.

Host Boot IPL

In POWER8 the initialization process during IPL has been changed a little bit. The Flexible Service Processor (FSP) is no longer the only instance that initializes and runs the bootup process. With POWER8 the FSP initializes the boot processes, but on the POWER8 Processors itself one part of the Firmware is running and performing the CEC chip initialization. A new component called the PNOR Chip stores the Host Boot firmware and the Self Boot Engine (SBE) is an internal part of the POWER8 Chip itself and is used to boot the chip.

With this Host Boot initialization new Progress codes are available. An example for a FSP Progress code is C1009003. During the Host Boot IPL Progress codes like CC009344 appear.

In case of a failure during the Host Boot process a new Host Boot System Dump is collected and stored. This type of dump includes Host Boot memory and will be offloaded to HMC when available.

Run time

All Power Systems servers can monitor critical system components during run time, and they can take corrective actions when recoverable faults occur. IBM hardware error-check architecture provides the ability to report non-critical errors in an *out-of-band* communications path to the service processor without affecting system performance.

A significant part of IBM runtime diagnostic capabilities originate with the service processor. Extensive diagnostic and fault analysis routines were developed and improved over many generations of POWER processor-based servers, and enable quick and accurate predefined responses to both actual and potential system problems.

The service processor correlates and processes runtime error information by using logic derived from IBM engineering expertise to count recoverable errors (called thresholding) and predict when corrective actions must be automatically initiated by the system. These actions can include the following items:

- Requests for a part to be replaced
- Dynamic invocation of built-in redundancy for automatic replacement of a failing part
- Dynamic deallocation of failing components so that system availability is maintained

Device drivers

In certain cases, diagnostics are best performed by operating system-specific drivers, most notably I/O devices that are owned directly by a logical partition. In these cases, the operating system device driver often works in conjunction with I/O device microcode to isolate and recover from problems. Potential problems are reported to an operating system device driver,

which logs the error. I/O devices can also include specific exercisers that can be invoked by the diagnostic facilities for problem recreation if required by service procedures.

4.3.3 Reporting

In the unlikely event that a system hardware or environmentally induced failure is diagnosed, IBM Power Systems servers report the error through various mechanisms. The analysis result is stored in system NVRAM. Error log analysis (ELA) can be used to display the failure cause and the physical location of the failing hardware.

With the integrated service processor, the system can automatically send an alert through a phone line to a pager, or call for service in the event of a critical system failure. A hardware fault also illuminates the amber system fault LED, located on the system unit, to alert the user of an internal hardware problem.

On POWER8 processor-based servers, hardware and software failures are recorded in the system log. When a management console is attached, an ELA routine analyzes the error, forwards the event to the Service Focal Point (SFP) application running on the management console, and has the capability to notify the system administrator that it has isolated a likely cause of the system problem. The service processor event log also records unrecoverable checkstop conditions, forwards them to the SFP application, and notifies the system administrator. After the information is logged in the SFP application, if the system is properly configured, a call-home service request is initiated and the pertinent failure data with service parts information and part locations is sent to the IBM service organization.This information will also contain the client contact information as defined in the IBM Electronic Service Agent (ESA) guided setup wizard. With the new HMC V8R8.1.0 a Serviceable Event Manager is available to manually block problems from being automatically transferred to IBM. Please refer to “Service Event Manager” on page 137 for more details.

Error logging and analysis

When the root cause of an error is identified by a fault isolation component, an error log entry is created with basic data such as the following examples:

- An error code that uniquely describes the error event
- The location of the failing component
- The part number of the component to be replaced, including pertinent data such as engineering and manufacturing levels
- Return codes
- Resource identifiers
- FFDC data

Data that contains information about the effect that the repair will have on the system is also included. Error log routines in the operating system and FSP can then use this information and decide whether the fault is a call-home candidate. If the fault requires support intervention, a call is placed with service and support, and a notification is sent to the contact that is defined in the ESA-guided setup wizard.

Remote support

The Remote Management and Control (RMC) subsystem is delivered as part of the base operating system, including the operating system that runs on the Hardware Management Console. RMC provides a secure transport mechanism across the LAN interface between the operating system and the Hardware Management Console and is used by the operating

system diagnostic application for transmitting error information. It performs several other functions also, but these are not used for the service infrastructure.

Service Focal Point (SFP)

A critical requirement in a logically partitioned environment is to ensure that errors are not lost before being reported for service, and that an error should be reported only once, regardless of how many logical partitions experience the potential effect of the error. The Manage Serviceable Events task on the management console is responsible for aggregating duplicate error reports, and ensures that all errors are recorded for review and management.

When a local or globally reported service request is made to the operating system, the operating system diagnostic subsystem uses the Remote Management and Control subsystem to relay error information to the Hardware Management Console. For global events (platform unrecoverable errors, for example), the service processor also forwards error notification of these events to the Hardware Management Console, providing a redundant error-reporting path in case of errors in the Remote Management and Control subsystem network.

The first occurrence of each failure type is recorded in the Manage Serviceable Events task on the management console. This task then filters and maintains a history of duplicate reports from other logical partitions on the service processor. It then looks at all active service event requests, analyzes the failure to ascertain the root cause and, if enabled, initiates a call home for service. This methodology ensures that all platform errors will be reported through at least one functional path, ultimately resulting in a single notification for a single problem.

Extended error data

Extended error data (EED) is additional data that is collected either automatically at the time of a failure or manually at a later time. The data that is collected depends on the invocation method but includes information such as firmware levels, operating system levels, additional fault isolation register values, recoverable error threshold register values, system status, and any other pertinent data.

The data is formatted and prepared for transmission back to IBM either to assist the service support organization with preparing a service action plan for the service representative or for additional analysis.

System-dump handling

In certain circumstances, an error might require a dump to be automatically or manually created. In this event, it is off-loaded to the management console. Specific management console information is included as part of the information that can optionally be sent to IBM support for analysis. If additional information that relates to the dump is required, or if viewing the dump remotely becomes necessary, the management console dump record notifies the IBM support center regarding on which management console the dump is located.

4.3.4 Notifying

After a Power Systems server detects, diagnoses, and reports an error to an appropriate aggregation point, it then takes steps to notify the client, and if necessary the IBM support organization. Depending on the assessed severity of the error and support agreement, this client notification might range from a simple notification to having field service personnel automatically dispatched to the client site with the correct replacement part.

Client Notify

When an event is important enough to report, but does not indicate the need for a repair action or the need to call home to IBM service and support, it is classified as *Client Notify*. Clients are notified because these events might be of interest to an administrator. The event might be a symptom of an expected systemic change, such as a network reconfiguration or failover testing of redundant power or cooling systems. These events include the following examples:

- Network events such as the loss of contact over a local area network (LAN)
- Environmental events such as ambient temperature warnings
- Events that need further examination by the client (although these events do not necessarily require a part replacement or repair action)

Client Notify events are serviceable events, by definition, because they indicate that something happened that requires client awareness if the client wants to take further action. These events can always be reported back to IBM at the discretion of the client.

Call home

Call home refers to an automatic or manual call from a customer location to an IBM support structure with error log data, server status, or other service-related information. The call home feature invokes the service organization so that the appropriate service action can begin. Call home can be done through HMC or most non-HMC managed systems. Although configuring a call home function is optional, clients are encouraged to implement this feature to obtain service enhancements such as reduced problem determination and faster and potentially more accurate transmittal of error information. In general, using the call home feature can result in increased system availability. The Electronic Service Agent application can be configured for automated call home. See 4.4.4, “Electronic Services and Electronic Service Agent” on page 135 for specific details.

Vital product data and inventory management

Power Systems store vital product data (VPD) internally, which keeps a record of how much memory is installed, how many processors are installed, the manufacturing level of the parts, and so on. These records provide valuable information that can be used by remote support and service representatives, enabling the representatives to provide assistance in keeping the firmware and software current on the server.

IBM problem management database

At the IBM support center, historical problem data is entered into the IBM Service and Support Problem Management database. All of the information that is related to the error, along with any service actions taken by the service representative, is recorded for problem management by the support and development organizations. The problem is then tracked and monitored until the system fault is repaired.

4.3.5 Locating and servicing

The final component of a comprehensive design for serviceability is the ability to effectively locate and replace parts requiring service. POWER processor-based systems use a combination of visual cues and guided maintenance procedures to ensure that the identified part is replaced correctly, every time.

Packaging for service

The following service enhancements are included in the physical packaging of the systems to facilitate service:

- Color coding (touch points)
  - Terra-cotta-colored touch points indicate that a component (FRU or CRU) can be concurrently maintained.
  - Blue-colored touch points delineate components that are not concurrently maintained (those that require the system to be turned off for removal or repair).
- Tool-less design

  Selected IBM systems support tool-less or simple tool designs. These designs require no tools, or require basic tools such as flathead screw drivers, to service the hardware components.
- Positive retention

  Positive retention mechanisms help to ensure proper connections between hardware components, such as from cables to connectors, and between two cards that attach to each other. Without positive retention, hardware components risk becoming loose during shipping or installation, preventing a good electrical connection. Positive retention mechanisms such as latches, levers, thumb-screws, pop Nylatches (U-clips), and cables are included to help prevent loose connections and aid in installing (seating) parts correctly. These positive retention items do not require tools.

Light Path

The Light Path LED feature is for low-end systems, including Power Systems through models Power S822, that can be repaired by clients. In the Light Path LED implementation, when a fault condition is detected on the POWER8 processor-based system, an amber FRU fault LED is illuminated, which is then rolled up to the system fault LED. The Light Path system pinpoints the exact part by lighting the amber FRU fault LED that is associated with the part to be replaced.

The system can clearly identify components for replacement by using specific component level LEDs, and can also guide the servicer directly to the component by signaling (remaining on, or *solid*) the system fault LED, enclosure fault LED, and the component FRU fault LED.

After the repair, the LEDs shut off automatically when the problem is fixed. The Light Path LEDs are only visible while system is in standby power. There is no gold cap or battery implemented.

Service labels

Service providers use these labels to assist in doing maintenance actions. Service labels are in various formats and positions, and are intended to transmit readily available information to the servicer during the repair process.

Several of these service labels and their purposes are described in the following list:

- Location diagrams are strategically positioned on the system hardware, relating information regarding the placement of hardware components. Location diagrams can include location codes, drawings of physical locations, concurrent maintenance status, or other data that is pertinent to a repair. Location diagrams are especially useful when multiple components are installed, such as DIMMs, sockets, processor cards, fans, adapter cards, LEDs, and power supplies.
- Remove or replace procedure labels contain procedures often found on a cover of the system or in other locations that are accessible to the servicer. These labels provide systematic procedures, including diagrams, detailing how to remove and replace certain serviceable hardware components.
- Numbered arrows are used to indicate the order of operation and serviceability direction of components. Various serviceable parts such as latches, levers, and touch points must be pulled or pushed in a certain direction and order so that the mechanical mechanisms can engage or disengage. Arrows generally improve the ease of serviceability.

The operator panel

The operator panel on a POWER processor-based system is an LCD display (four rows by sixteen elements) that is used to present boot progress codes, indicating advancement through the system power-on and initialization processes. The operator panel is also used to display error and location codes when an error occurs that prevents the system from booting. It includes several buttons, enabling a service support representative (SSR) or client to change various boot-time options and for other limited service functions.

Concurrent maintenance

The IBM POWER8 processor-based systems are designed with the understanding that certain components have higher intrinsic failure rates than others. The movement of fans, power supplies, and physical storage devices naturally make them more susceptible to wearing down or burning out. Other devices, such as I/O adapters can begin to wear from repeated plugging and unplugging. For these reasons, these devices are specifically designed to be concurrently maintainable when properly configured.

In other cases, a client might be in the process of moving or redesigning a data center or planning a major upgrade. At those times, flexibility is crucial. The IBM POWER8 processor-based systems are designed for redundant or concurrently maintainable power, fans and physical storage.

The most recent members of the IBM Power Systems family, based on the POWER8 processor, continue to support concurrent maintenance of power, cooling, PCI adapters, media devices, and the operator panel. The operator Panel can be replaced concurrently from the ASMI menu. In addition, they support concurrent firmware fix pack updates when possible. The determination of whether a firmware fix pack release can be updated concurrently is identified in the readme file that is released with the firmware.

Repair and verify system

Repair and verify (R&V) is a system that is used to guide a service provider, step-by-step, through the process of repairing a system and verifying that the problem was repaired. The steps are customized in the appropriate sequence for the particular repair for the specific system being repaired. The following scenarios are covered by repair and verify:

- Replacing a defective field-replaceable unit (FRU) or a customer-replaceable unit (CRU)
- Reattaching a loose or disconnected component
- Correcting a configuration error
- Removing or replacing an incompatible FRU
- Updating firmware, device drivers, operating systems, middleware components, and IBM applications after replacing a part

Repair and verify procedures can be used by service representative providers who are familiar with the task and those who are not. Education-on-demand content is placed in the procedure at the appropriate locations. Throughout the repair and verify procedure, repair history is collected and provided to the Service and Support Problem Management Database for storage with the serviceable event, to ensure that the guided maintenance procedures are operating correctly.

If a server is managed by a management console, then many of the repair and verify procedures are done from the management console. If the FRU to be replaced is a PCI adapter or an internal storage device, the service action is always performed from the operating system of the partition owning that resource.

Clients can subscribe through the subscription services to obtain the notifications about the latest updates available for service-related documentation. The latest version of the documentation is accessible through the Internet.

4.4 Manageability

Several functions and tools help manageability so you can efficiently and effectively manage your system.

4.4.1 Service user interfaces

The service interface allows support personnel or the client to communicate with the service support applications in a server by using a console, interface, or terminal. Delivering a clear, concise view of available service applications, the service interface allows the support team to manage system resources and service information in an efficient and effective way. Applications that are available through the service interface are carefully configured and placed to give service providers access to important service functions.

Various service interfaces are used, depending on the state of the system and its operating environment. The primary service interfaces are the following items:

- Light Path

  See “Light Path” on page 125 and “Service labels” on page 126.
- Service processor, Advanced System Management Interface (ASMI)
- Operator panel
- Operating system service menu
- Service Focal Point on the Hardware Management Console
- Service Focal Point Lite on Integrated Virtualization Manager

Service processor

The service processor is a controller that is running its own operating system. It is a component of the service interface card.

The service processor operating system has specific programs and device drivers for the service processor hardware. The host interface is a processor support interface that is connected to the POWER processor. The service processor is always working, regardless of the main system unit's state. The system unit can be in the following states:

- Standby (power off)
- Operating, ready to start partitions
- Operating with running logical partitions

The service processor is used to monitor and manage the system hardware resources and devices. The service processor checks the system for errors, ensuring that the connection to the management console for manageability purposes and accepting Advanced System Management Interface (ASMI) Secure Sockets Layer (SSL) network connections. The service processor provides the ability to view and manage the machine-wide settings by using the ASMI, and enables complete system and partition management from the HMC.

Analyze system that does not boot: The flexible service processor enables a system that does not boot to be analyzed. Reference codes and detailed data is available in the ASMI and will be transferred to the HMC.

The service processor uses two Ethernet ports that run at 1 Gbps speed. Consider the following information:

- Both Ethernet ports are visible only to the service processor and can be used to attach the server to an HMC or to access the ASMI. The ASMI options can be accessed through an HTTP server that is integrated into the service processor operating environment.
- Both Ethernet ports support only auto-negotiation. Customer-selectable media speed and duplex settings are not available.
- Both Ethernet ports have a default IP address, as follows:
  - Service processor eth0 (HMC1 port) is configured as 169.254.2.147.
  - Service processor eth1 (HMC2 port) is configured as 169.254.3.147.

The following functions are available through service processor:

- Call home
- Advanced System Management Interface (ASMI)
- Error Information (error code, part number, location codes) menu
- View of guarded components
- Limited repair procedures
- Generate dump
- LED Management menu
- Remote view of ASMI menus
- Firmware update through USB key

Advanced System Management Interface

Advanced System Management Interface (ASMI) is the interface to the service processor that enables you to manage the operation of the server, such as auto-power restart, and to view information about the server, such as the error log and vital product data. Various repair procedures require connection to the ASMI.

The ASMI is accessible through the management console. It is also accessible by using a web browser on a system that is connected directly to the service processor (in this case, either a standard Ethernet cable or a crossed cable) or through an Ethernet network. ASMI can also be accessed from an ASCII terminal, but this is only available while the system is in the platform powered-off mode.

Use the ASMI to change the service processor IP addresses or to apply certain security policies and prevent access from undesired IP addresses or ranges.

You might be able to use the service processor's default settings. In that case, accessing the ASMI is not necessary. To access ASMI, use one of the following methods:

- Use a management console.

  If configured to do so, the management console connects directly to the ASMI for a selected system from this task.

  To connect to the Advanced System Management interface from a management console, use the following steps:

  a. Open **Systems Management** from the navigation pane.

  b. From the work panel, select one managed systems to work with.

  c. From the System Management tasks list, select **Operations > Launch Advanced System Management** (ASM).

- Use a web browser.

  At the time of writing, supported web browsers are Microsoft Internet Explorer (Version 10.0.9200.16439), Mozilla Firefox (Version 24), and Chrome (Version 30). Later versions of these browsers might work but are not officially supported. The JavaScript language and cookies must be enabled and TLS 1.2 might have to be enabled.

  The web interface is available during all phases of system operation, including the initial program load (IPL) and run time. However, several of the menu options in the web interface are unavailable during IPL or run time to prevent usage or ownership conflicts if the system resources are in use during that phase. The ASMI provides a Secure Sockets Layer (SSL) web connection to the service processor. To establish an SSL connection, open your browser by using the following address:

  `https://<ip_address_of_service_processor>`

  Note: To make the connection through Internet Explorer, click **Tools Internet Options**. Clear the **Use TLS 1.0** check box, and click **OK**.

- Use an ASCII terminal.

  The ASMI on an ASCII terminal supports a subset of the functions that are provided by the web interface and is available only when the system is in the platform powered-off mode. The ASMI on an ASCII console is not available during several phases of system operation, such as the IPL and run time.

- Command line start of the ASMI

  Either on the HMC itself or when properly configured on a remote system it is possible to start ASMI Web interface also from the HMC commandline. Open a Terminal window on the HMC or access the HMC with a Terminal emulation and issue:

  ```
  asmmenu --ip <ip address>
  ```

  On the HMC itself a Browser window will open automatically with the ASMI screen and when configured properly also a Browser window will open on a remote system when issued from there.

The operator panel

The service processor provides an interface to the operator panel, which is used to display system status and diagnostic information.

The operator panel can be accessed in two ways:

- By using the normal operational front view.
- By pulling it out to access the switches and viewing the LCD display.

Several of the operator panel features include the following items:

- A 2 x 16 character LCD display
- Reset, enter, power On/Off, increment, and decrement buttons
- Amber System Information/Attention, green Power LED
- Blue Enclosure Identify LED on the Power S822
- Altitude sensor
- USB Port
- Speaker/Beeper

The following functions are available through the operator panel:

- Error Information
- Generate dump
- View machine type, model, and serial number
- Limited set of repair functions

Operating system service menu

The system diagnostics consist of stand-alone diagnostics that are loaded from the DVD drive, and online diagnostics (available in AIX).

Online diagnostics, when installed, are a part of the AIX operating system on the disk or server. They can be booted in single-user mode (service mode), run in maintenance mode, or run concurrently (concurrent mode) with other applications. They have access to the AIX error log and the AIX configuration data.

The modes are as follows:

- Service mode

  This mode requires a service mode boot of the system and enables the checking of system devices and features. Service mode provides the most complete self-check of the system resources. All system resources, except the SCSI adapter and the disk drives used for paging, can be tested.

- Concurrent mode

  This mode enables the normal system functions to continue while selected resources are being checked. Because the system is running in normal operation, certain devices might require additional actions by the user or diagnostic application before testing can be done.

- Maintenance mode

  This mode enables the checking of most system resources. Maintenance mode provides the same test coverage as service mode. The difference between the two modes is the way that they are invoked. Maintenance mode requires that all activity on the operating system be stopped. The **shutdown -m** command is used to stop all activity on the operating system and put the operating system into maintenance mode.

The System Management Services (SMS) error log is accessible on the SMS menus. This error log contains errors that are found by partition firmware when the system or partition is booting.

The service processor's error log can be accessed on the ASMI menus.

You can also access the system diagnostics from a Network Installation Management (NIM) server.

Alternate method: When you order a Power System, a DVD-ROM or DVD-RAM might be optional. An alternate method for maintaining and servicing the system must be available if you do not order the DVD-ROM or DVD-RAM.

Depending on the operating system, the following service-level functions are what you typically see when you use the operating system service menus:

- Product activity log
- Trace Licensed Internal Code
- Work with communications trace
- Display/Alter/Dump
- Licensed Internal Code log
- Main storage dump manager
- Hardware service manager
- Call Home/Customer Notification
- Error information menu
- LED management menu
- Concurrent/Non-concurrent maintenance (within scope of the OS)
- Managing firmware levels
  - Server
  - Adapter
- Remote support (access varies by OS)

Service Focal Point on the Hardware Management Console

Service strategies become more complicated in a partitioned environment. The Manage Serviceable Events task in the management console can help to streamline this process.

Each logical partition reports errors that it detects and forwards the event to the Service Focal Point (SFP) application that is running on the management console, without determining whether other logical partitions also detect and report the errors. For example, if one logical partition reports an error for a shared resource, such as a managed system power supply, other active logical partitions might report the same error.

By using the *Manage Serviceable Events* task in the management console, you can avoid long lists of repetitive call-home information by recognizing that these are repeated errors and consolidating them into one error.

In addition, you can use the Manage Serviceable Events task to initiate service functions on systems and logical partitions, including the exchanging of parts, configuring connectivity, and managing dumps.

4.4.2 IBM Power Systems firmware maintenance

The IBM Power Systems Client-Managed Microcode is a methodology that enables you to manage and install microcode updates on Power Systems and associated I/O adapters.

Firmware entitlement

With the new HMC Version V8R8.1.0.0 and POWER Servers the firmware installations will be restricted to entitled servers. The customer must be registered with IBM and entitled with a

service contract. During the initial machine warranty period the access key is already installed in the machine from manufacturing. The key will be valid for the regular warranty period plus some additional time. The POWER System firmware is being relocated from the public repository to the access control repository. I/O firmware will remain on the public repository but server must be entitled for installation. When the **lslic** command is used to display the Firmware Levels a new value '**update_access_key_exp_date**' is added. The HMC GUI as well as the ASMI menu show the Update access key expiration date.

When the system is no longer entitled the Firmware updates will fail. Some new SRCs are available:

- E302FA06: Acquisition entitlement check failed
- E302FA08: Installation entitlement check failed

Any Firmware release that was made available during the entitled time frame can still be installed. For example, the entitlement period ends on December 31st 2014 and a new Firmware release is release prior that entitlement period, then it can still be installed. Even when that Firmware is downloaded after December 31st 2014, but it still was made available prior the end of the entitlement period it can be installed. Any newer release requires a new update access key.

Note: The update access key expiration date requires a valid entitlement of the system to perform Firmware updates.

This key will be available on the IBM CoD Home page:

http://www.ibm.com/pod/pod

To access the IBM entitled Software Support page for further details, access:

http://www.ibm.com/servers/eserver/ess

Firmware updates

System firmware is delivered as a release level or a service pack. Release levels support the general availability (GA) of new functions or features, and new machine types or models. Upgrading to a higher release level is disruptive to customer operations. IBM intends to introduce no more than two new release levels per year. These release levels will be supported by service packs. Service packs are intended to contain only firmware fixes and not to introduce new function. A *service pack* is an update to an existing release level.

If the system is managed by a management console, you use the management console for firmware updates. By using the management console, you can take advantage of the Concurrent Firmware Maintenance (CFM) option when concurrent service packs are available. CFM is the IBM term used to describe the IBM Power Systems firmware updates that can be partially or wholly concurrent or non disruptive. With the introduction of CFM, IBM is significantly increasing a client's opportunity to stay on a given release level for longer periods of time. Clients that want maximum stability can defer until there is a compelling reason to upgrade, such as the following reasons:

- A release level is approaching its end-of-service date (that is, it has been available for about a year, and soon, service will not be supported).
- Move a system to a more standardized release level when there are multiple systems in an environment with similar hardware.
- A new release has new functionality that is needed in the environment.

- A scheduled maintenance action will cause a platform reboot, which provides an opportunity to also upgrade to a new firmware release.

The updating and upgrading of system firmware depends on several factors, such as whether the system is stand-alone or managed by a management console, the current firmware installed, and what operating systems are running on the system. These scenarios and the associated installation instructions are comprehensively outlined in the firmware section of Fix Central:

http://www.ibm.com/support/fixcentral/

You might also want to review the best practice white papers:

http://www14.software.ibm.com/webapp/set2/sas/f/best/home.html

Firmware update steps

The system firmware consists of service processor microcode, Open Firmware microcode, SPCN microcode.

The firmware and microcode can be downloaded and installed either from an HMC, from a running partition, or from USB port number 1 on the rear, if that system is not managed by an HMC.

Power Systems has a permanent firmware boot side (A side) and a temporary firmware boot side (B side). New levels of firmware must be installed first on the temporary side to test the update's compatibility with existing applications. When the new level of firmware is approved, it can be copied to the permanent side.

For access to the initial web pages that address this capability, see the Support for IBM Systems web page:

http://www.ibm.com/systems/support

For Power Systems, select the **Power** link.

Although the content under the Popular links section can change, click the **Firmware and HMC updates** link to go to the resources for keeping your system's firmware current.

If there is an HMC to manage the server, the HMC interface can be use to view the levels of server firmware and power subsystem firmware that are installed and that are available to download and install.

Each IBM Power Systems server has the following levels of server firmware and power subsystem firmware:

- Installed level

  This level of server firmware or power subsystem firmware is installed and will be installed into memory after the managed system is powered off and then powered on. It is installed on the temporary side of system firmware.

- Activated level

  This level of server firmware or power subsystem firmware is active and running in memory.

- Accepted level

  This level is the backup level of server or power subsystem firmware. You can return to this level of server or power subsystem firmware if you decide to remove the installed level. It is installed on the permanent side of system firmware.

In Figure 4-2 the HMC screen shows the different levels.

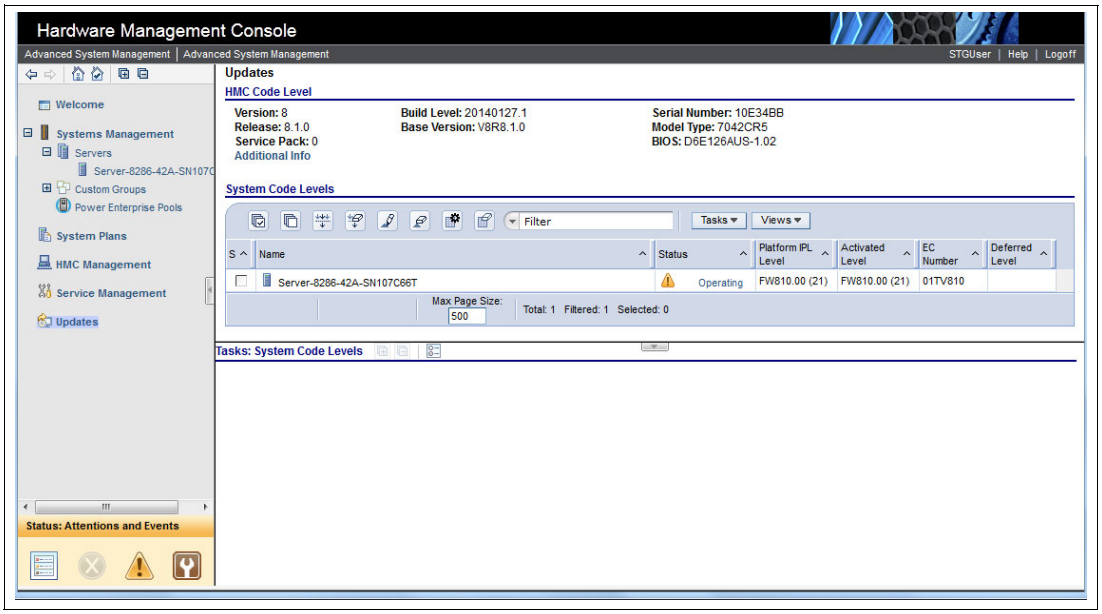


Figure 4-2 HMC System Firmware Screen

IBM provides the Concurrent Firmware Maintenance (CFM) function on selected Power Systems. This function supports applying non disruptive system firmware service packs to the system concurrently (without requiring a reboot operation to activate changes). For systems that are not managed by an HMC, the installation of system firmware is always disruptive.

The concurrent levels of system firmware can, on occasion, contain fixes that are known as *deferred*. These deferred fixes can be installed concurrently but are not activated until the next IPL. Deferred fixes, if any, will be identified in the Firmware Update Descriptions table of the firmware document. For deferred fixes within a service pack, only the fixes in the service pack that cannot be concurrently activated are deferred. Table 4-1 shows the file-naming convention for system firmware.

Table 4-1 Firmware naming convention

PPNNSSS_FFF_DDD			
PP	Package identifier	01	-

PPNNSSS_FFF_DDD			
NN	Platform and class	SV	Low end
SSS	Release indicator		
FFF	Current fix pack		
DDD	Last disruptive fix pack		

The following example uses the convention:

```
01SV810_030_030 = POWER8 Entry Systems Firmware for 8284-22A
```

An installation is disruptive if the following statements are true:

- The release levels (SSS) of currently installed and new firmware differ.
- The service pack level (FFF) and the last disruptive service pack level (DDD) are equal in new firmware.

Otherwise, an installation is concurrent if the service pack level (FFF) of the new firmware is higher than the service pack level currently installed on the system and the conditions for disruptive installation are not met.

4.4.3 Concurrent firmware update improvements

Since POWER6, firmware service packs are generally concurrently applied and take effect immediately. Occasionally, a service pack is shipped where most of the features can be concurrently applied; but because changes to some server functions (for example, changing initialization values for chip controls) cannot occur during operation, a patch in this area required a system reboot for activation.

With the Power-On Reset Engine (PORE), the firmware can now dynamically power off processor components, make changes to the registers and re-initialize while the system is running, without discernible impact to any applications running on a processor. This potentially allows concurrent firmware changes in POWER8, which in earlier designs, required a reboot to take effect.

Activating some new firmware functions requires installation of a firmware release level. This process is disruptive to server operations and requires a scheduled outage and full server reboot.

4.4.4 Electronic Services and Electronic Service Agent

IBM transformed its delivery of hardware and software support services to help you achieve higher system availability. Electronic Services is a web-enabled solution that offers an exclusive, no-additional-charge enhancement to the service and support that is available for IBM servers. These services provide the opportunity for greater system availability with faster problem resolution and preemptive monitoring. The Electronic Services solution consists of two separate, but complementary, elements:

Electronic Services news page

The Electronic Services news page is a single Internet entry point that replaces the multiple entry points, which are traditionally used to access IBM Internet services and support. With the news page, you can gain easier access to IBM resources for assistance in resolving technical problems.

Electronic Service Agent

The Electronic Service Agent is software that resides on your server. It monitors events and transmits system inventory information to IBM on a periodic, client-defined timetable. The Electronic Service Agent automatically reports hardware problems to IBM.

Early knowledge about potential problems enables IBM to deliver proactive service that can result in higher system availability and performance. In addition, information that is collected through the Service Agent is made available to IBM service support representatives when they help answer your questions or diagnose problems. Installation and use of IBM Electronic Service Agent for problem reporting enables IBM to provide better support and service for your IBM server.

To learn how Electronic Services can work for you, visit the following site; an IBM ID is required:

http://www.ibm.com/support/electronic

Benefits are as follows:

- Increased uptime

  The Electronic Service Agent tool is designed to enhance the warranty or maintenance agreement by providing faster hardware error reporting and uploading system information to IBM Support. This way can translate to less wasted time monitoring the symptoms, diagnosing the error, and manually calling IBM Support to open a problem record.

  Its 24x7 monitoring and reporting mean no more dependence on human intervention or off-hours customer personnel when errors are encountered in the middle of the night.

- Security

  The Electronic Service Agent tool is designed to be secure in monitoring, reporting, and storing the data at IBM. The Electronic Service Agent tool securely transmits either with the Internet (HTTPS or VPN) or modem, and can be configured to communicate securely through gateways to provide customers a single point of exit from their site.

  Communication is one way. Activating Electronic Service Agent does not enable IBM to call into a customer's system. System inventory information is stored in a secure database, which is protected behind IBM firewalls. It is viewable only by the customer and IBM. The customer's business applications or business data is never transmitted to IBM.

- More accurate reporting

  Because system information and error logs are automatically uploaded to the IBM Support center in conjunction with the service request, customers are not required to find and send system information, decreasing the risk of misreported or misdiagnosed errors.

  When inside IBM, problem error data is run through a data knowledge management system and knowledge articles are appended to the problem record.

- Customized support

  By using the IBM ID that you enter during activation, you can view system and support information by selecting **My Systems** at the Electronic Support website:

  http://www.ibm.com/support/electronic

  My Systems provides valuable reports of installed hardware and software, using information collected from the systems by Electronic Service Agent. Reports are available for any system associated with the customers IBM ID. Premium Search combines the function of search and the value of Electronic Service Agent information, providing advanced search of the technical support knowledge base. Using Premium Search and the Electronic Service Agent information that was collected from your system, your clients are able to see search results that apply specifically to their systems.

For more information about how to use the power of IBM Electronic Services, contact your IBM Systems Services Representative, or visit the following website:

http://www.ibm.com/support/electronic

Service Event Manager

Allows the user to decide which of the Serviceable Events will be called home with the Electronic Service Agent or not. It is possible to lock certain Events. Some customers might not allow data to be transferred outside their company. Once the SEM is enabled the analysis of the possible problems might take longer.

- The SEM can be enabled with the following syntax:

  chhmc -c sem -s enable

- Disabling SEM mode and specify what state to leave call home in.

  chhmc -c sem -s disable --callhome disable

  chhmc -c sem -s disable --callhome enable

From the HMC GUI the basic configuration of the SEM can be performed. After selecting the Service Event Manager as shown in Figure 4-3 the HMC console has to be added.

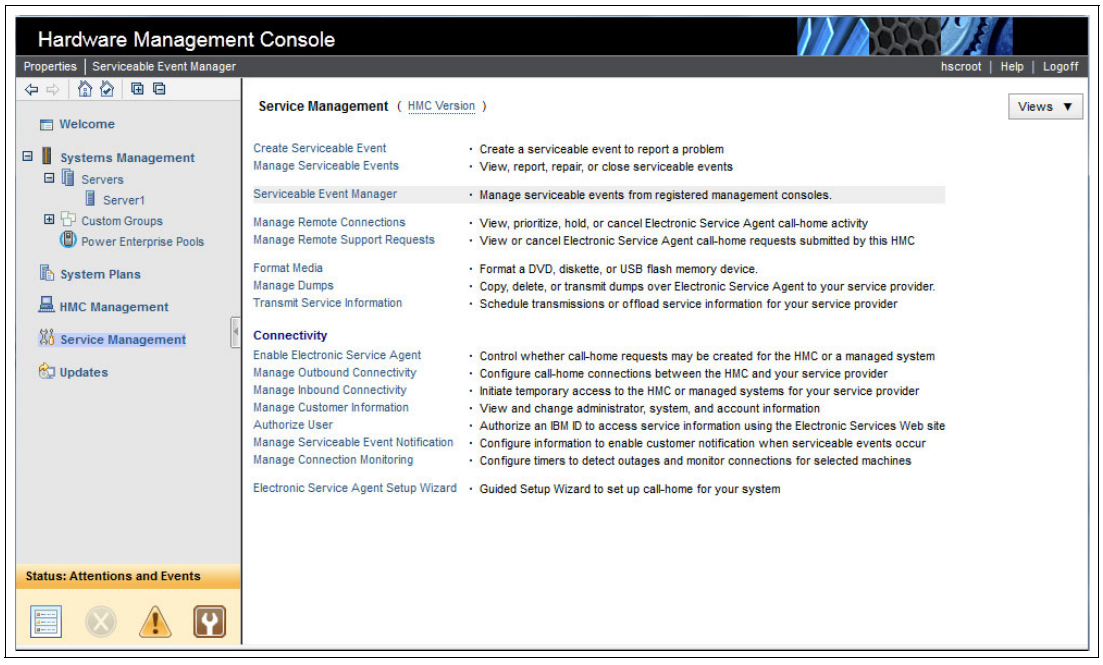

Hardware Management Console
Properties | Serviceable Event Manager
hscroot | Help | Logoff
Welcome
Systems Management
Servers
Server1
Custom Groups
Power Enterprise Pools
System Plans
HMC Management
Service Management
Updates
Status: Attentions and Events
Service Management (HMC Version)
Views
Create Serviceable Event
Create a serviceable event to report a problem
Manage Serviceable Events
View, report, repair, or close serviceable events
Serviceable Event Manager
Manage serviceable events from registered management consoles.
Manage Remote Connections
View, prioritize, hold, or cancel Electronic Service Agent call-home activity
Manage Remote Support Requests
View or cancel Electronic Service Agent call-home requests submitted by this HMC
Format Media
Format a DVD, diskette, or USB flash memory device.
Manage Dumps
Copy, delete, or transmit dumps over Electronic Service Agent to your service provider.
Transmit Service Information
Schedule transmissions or offload service information for your service provider
Connectivity
Enable Electronic Service Agent
Control whether call-home requests may be created for the HMC or a managed system
Manage Outbound Connectivity
Configure call-home connections between the HMC and your service provider
Manage Inbound Connectivity
Initiate temporary access to the HMC or managed systems for your service provider
Manage Customer Information
View and change administrator, system, and account information
Authorize User
Authorize an IBM ID to access service information using the Electronic Services Web site
Manage Serviceable Event Notification
Configure information to enable customer notification when serviceable events occur
Manage Connection Monitoring
Configure timers to detect outages and monitor connections for selected machines
Electronic Service Agent Setup Wizard
Guided Setup Wizard to set up call-home for your system

Figure 4-3 HMC selection for Service Event Manager

The next window allows the user to configure the HMC that is used to manage the Serviceable Events and proceed with further configuration steps. See Figure 4-4 for the initial screen:

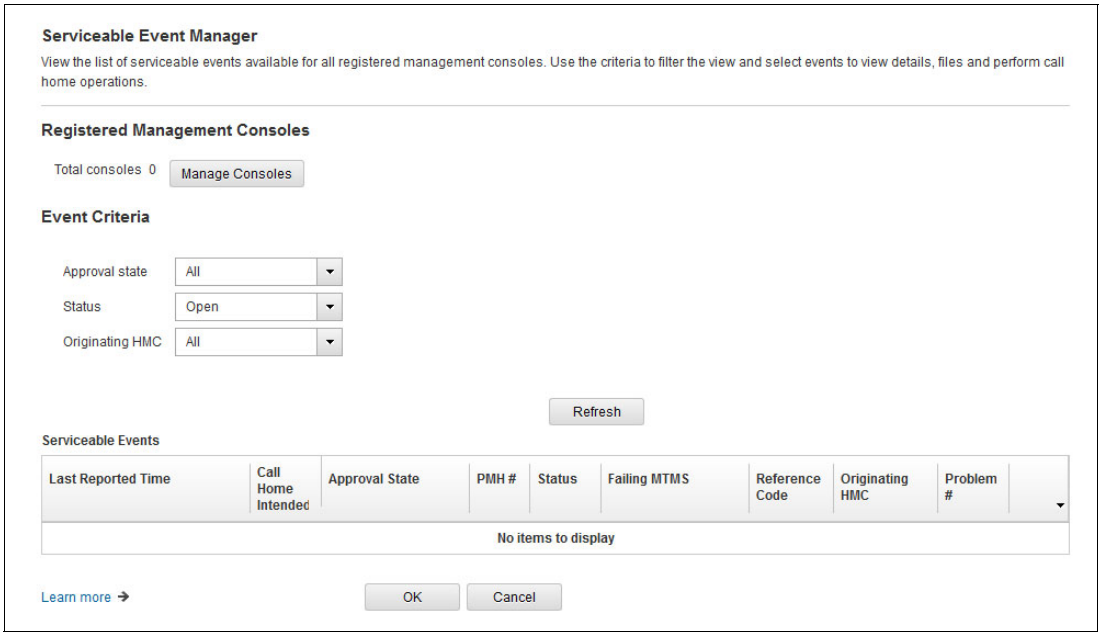


Figure 4-4 Initial SEM screen

Detailed description of the different configurable options:

- Registered Management Consoles

  Total consoles lists the number of consoles registered.

  Select Manage Consoles to mange the list of registered management consoles.
- Event Criteria

  Select the filters for filtering the list of serviceable events shown.

  Once selections are made, select Refresh to refresh the list based on the filter values.
- Approval state

  Select the value for approval state to filter the list.
- Status

  Select the value for status to filter the list.
- Originating HMC

  Select a single registered console or All consoles to filter the list.
- Serviceable Events

  The Serviceable Events table displays the list of events based on the filters selected.

  To refresh the list select Refresh.

The following context menu options are available when selecting an event in the table.

- View Details...

  Display the details of this event.

- View Files...

  Display the files associated with this event.

- Approve Call Home

  Approve the call home of this event. This option is only available if the event has not been approved already.

The Help / Learn more function can be used to get more details in the other available screens for the Serviceable Event Manager.

4.5 POWER8 RAS enhancements

POWER7 and POWER7+ systems reliability, availability and serviceability features are well documented in the following whitepaper:

http://public.dhe.ibm.com/common/ssi/ecm/en/pow03056usen/POW03056USEN.PDF

This document can still be used for an in-depth understanding of the available RAS Features in POWER7. The differences in the POWER8 server family are highlighted as follows:

- POWER8 Processor
  - The POWER8 processor module has a maximum of 12 cores compared to a maximum of 8 cores in POWER7
- On Chip Controller (OCC)
  - No need anymore for a separate module to handle Power Management and Thermal Monitoring.
  - The On Chip Controller (OCC) is integrated into each processor module. This OCC is separate from any customer accessible processor core. It is used to execute power management and thermal monitoring functions that used to be a function of a separate module in POWER7 (the TPMD). The OCC, in addition, can also be programmed execute other RAS related functions independent of any host processor.
- Integrated PCIe Controller
  - No more external I/O hub controller needed. Each processor module can directly drive two I/O slots or devices. Error handling and recovery remains similar to POWER7/POWER7+.
- Coherence Attach Processor Interface (CAPI)
- Fabric Bus Lane repair
  - similar to POWER7+
  - handles all external fabric busses
  - no need to use the Power On Reset Engine (PORE)
- Additional Internal Fabric Bus address error generation and checking
  - A more precise analysis of faults across the SMP interconnection
- Memory Control Replay buffer
  - Provides additional soft error protection in the memory buffer
- The Memory subsystem has an L4 Cache implemented
  - ECC protected
  - Data can be purged on errors similar to L2/L3 cache handling

– to prevent outages the L4 Cache has the capability to perform a bank delete so that faulty portions of the DRAM can be removed from the use

4.5.1 POWER7+ RAS features

Besides the improvements in the POWER8 Systems this section lists the existing POWER7+ RAS features that are in common with the new POWER8:

- Power-On Reset Engine (PORE)

  Enables a processor to be re-initialized while the system remains running. This feature will allow for the concurrent firmware updates situation, in which a processor initialization register value needs to be changed. Concurrent firmware updates might be more prevalent.

- L3 cache dynamic column repair

  This self-healing capability completes cache-line delete and uses the PORE feature to potentially avoid some repair actions or outages that are related to L3 cache.

- Accelerator RAS

  New accelerators are designed with RAS features to avoid system outages in the vast majority of faults that can be detected by the accelerators.

- Fabric Bus Dynamic Lane Repair

  POWER7+ has spare bit lanes that can dynamically be repaired (using PORE). This feature avoids any repair action or outage related to a single bit failure for the fabric bus.

4.6 Power-On Reset Engine

The POWER7+ chip includes a Power-On Reset Engine (PORE), a programmable hardware sequencer responsible for restoring the state of a powered down processor core and L2 cache (deep sleep mode), or chiplet (winkle mode). When a processor core wakes up from sleep or winkle, the PORE fetches code created by the POWER Hypervisor from a special location in memory containing the instructions and data necessary to restore the processor core to a functional state. This memory image includes all the necessary boot and runtime configuration data that were applied to this processor core since power-on, including circuit calibration and cache repair registers that are unique to each processor core. Effectively the PORE performs a mini initial program load (IPL) of the processor core or chiplet, completing the sequence of operations necessary to restart instruction execution, such as removing electrical and logical fences and re initializing the Digital PLL clock source.

Because of its special ability to perform clocks-off and clocks-on sequencing of the hardware, the PORE can also be used for RAS purposes:

- The service processor can use the PORE to concurrently apply an initialization update to a processor core/chiplet by loading new initialization values into memory and then forcing it to go in and out of winkle mode. This step happens, all without causing disruption to the workloads or operating system (all occurring in a few milliseconds).
- In the same fashion, PORE can initiate an L3 cache dynamic “bit-line” repair operation if the POWER Hypervisor detects too many recoverable errors in the cache.
- The PORE can be used to dynamically repair node-to-node fabric bit lanes in a POWER7+ processor-based server by quickly suspending chip-chip traffic during run time, reconfiguring the interface to use a spare bit lane, then resuming traffic, all without causing disruption to the operation of the server.

4.7 Operating system support for RAS features

Table 4-2 gives an overview of features for continuous availability that are supported by the various operating systems running on power systems. In the table, the word “Most” means most functions.

Table 4-2 Operating system support for RAS features

RAS feature	AIX 6.1	AIX 7.1	RHEL 6.5	SLES11 SP3
System deallocation of failing components				
Processor Fabric Bus Protection	X	X	X	X
Dynamic Processor Deallocation	X	X	X	X
Dynamic Processor Sparing	X	X	X	X
Processor Instruction Retry	X	X	X	X
Alternate Processor Recovery	X	X	X	X
Partition Contained Checkstop	X	X	X	X
Persistent processor deallocation	X	X	X	X
Optional ECC I/O hub with freeze behavior	X	X	X	X
PCI bus extended error detection	X	X	X	X
PCI bus extended error recovery	X	X	Most	Most
PCI-PCI bridge Enhanced Error Handling	X	X	-	-
PCI card hot-swap	X	X	X	X
Memory sparing with CoD at IPL time	X	X	X	X
Memory availability				
ECC memory, L2, L3 cache	X	X	X	X
CRC plus retry on memory data bus	X	X	X	X
Data Bus	X	X	X	X
Dynamic memory channel repair	X	X	X	X
Processor memory controller memory scrubbing	X	X	X	X
Memory page deallocation	X	X	X	X
Chipkill memory	X	X	X	X
L1 instruction and data array protection	X	X	X	X
L2/L3 ECC and cache line delete	X	X	X	X
Special uncorrectable error handling	X	X	X	X
Memory Buffer L4 Cache	X	X	X	X
Fault detection and isolation				
Platform FFDC diagnostics	X	X	X	X

RAS feature	AIX 6.1	AIX 7.1	RHEL 6.5	SLES11 SP3
Run-time diagnostics	X	X	Most	Most
Storage Protection Keys	X	X	-	-
Dynamic Trace	X	X	-	X
Operating System FFDC	X	X	-	-
Error log analysis	X	X	X	X
Freeze mode of I/O Hub	X	X	-	-
Service processor support for:				
► Built-in self-tests (BIST) for logic and arrays	X	X	X	X
► Wire tests	X	X	X	X
► Component initialization	X	X	X	X
Serviceability				
Boot-time progress indicators	X	X	Most	Most
Electronic Service Agent Call Home from management console	X	X	X	X
Firmware error codes	X	X	X	X
Operating system error codes	X	X	Most	Most
Inventory collection	X	X	X	X
Environmental and power warnings	X	X	X	X
Hot-plug fans, power supplies	X	X	X	X
Extended error data collection	X	X	X	X
SP mutual surveillance with POWER Hypervisor	X	X	X	X
Concurrent firmware update with Hardware management console	X	X	X	X
PORE: Core Inizialization without reboot	X	X	X	X
Service processor support for BIST	X	X	X	X
Electronic Service Agent Call Home Application	X	X	-	-
Guiding light LEDs	X	X	X	X
System dump for memory, SP	X	X	X	X
Information center / Systems Support Site service publications	X	X	X	X
System Support Site education	X	X	X	X
Operating system error reporting to management console SFP	X	X	X	X
RMC secure error transmission subsystem	X	X	X	X

RAS feature	AIX 6.1	AIX 7.1	RHEL 6.5	SLES11 SP3
Health check scheduled operations with management console	X	X	X	X
Operator panel (real or virtual)	X	X	X	X
Concurrent operator panel maintenance	X	X	X	X
Redundant Hardware management consoles	X	X	X	X
Automated server recovery/restart	X	X	X	X
PowerVM Live Partition Mobility	X	X	X	X
Live Application Mobility	X	X	-	-
Repair and Verify Guided Maintenance	X	X	Most	Most
Concurrent kernel update	X	X	X	X
Concurrent Hot Add/Repair Maintenance	X	X	X	X
Power and cooling				
Redundant, hot swap fans and blower for CEC	X	X	X	X
Redundant, hot swap power for CEC	X	X	X	X
CEC power/thermal sensor (CPU and memory)	X	X	X	X

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this paper.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- *Performance Optimization and Tuning Techniques for IBM Processors, including IBM POWER8,* SG24-8171
- *IBM Power Systems SR-IOV Technical Overview and Introduction,* REDP-5065
- *IBM Power System S812L and S822L Technical Overview and Introduction,* REDP-5098
- *IBM Power System S814 and S824 Technical Overview and Introduction,* REDP-5097
- *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940
- *IBM PowerVM Virtualization Managing and Monitoring*, SG24-7590
- *IBM PowerVM Best Practices*, SG24-8062
- *IBM PowerVM Enhancements What is new in 2013,* SG24-8198
- *IBM Power Systems HMC Implementation and Usage Guide*, SG24-7491

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

© Copyright IBM Corp. 2014. All rights reserved.

Other publications

These publications are also relevant as further information sources:

- IBM Power Facts and Features - IBM Power Systems, IBM PureFlex and Power Blades
  http://www.ibm.com/systems/power/hardware/reports/factsfeatures.html
- IBM Power S812L server specifications
  http://www.ibm.com/systems/power/hardware/s812l-s822l/specs.html
- IBM Power S822L server specifications
  http://www.ibm.com/systems/power/hardware/s812l-s822l/specs.html
- IBM Power S822 server specifications
  http://www.ibm.com/systems/power/hardware/s822/specs.html
- IBM Power S814 server specifications
  http://www.ibm.com/systems/power/hardware/s814/specs.html
- IBM Power S824 server specifications
  http://www.ibm.com/systems/power/hardware/s824/specs.html
- IBM EnergyScale for POWER7 Processor-Based Systems whitepaper
  http://public.dhe.ibm.com/common/ssi/ecm/en/pow03039usen/POW03039USEN.PDF
- Active Memory Expansion: Overview and Usage Guide
  http://public.dhe.ibm.com/common/ssi/ecm/en/pow03037usen/POW03037USEN.PDF
- Specific storage devices supported for Virtual I/O Server
  http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/datasheet.html
- System RAS - Introduction to Power Systems Reliability, Availability, and Serviceability
  http://public.dhe.ibm.com/common/ssi/ecm/en/pow03056usen/POW03056USEN.PDF

Online resources

These websites are also relevant as further information sources:

- IBM Knowledge Center

  http://www.ibm.com/support/knowledgecenter/
- IBM Power Systems Hardware Information Center

  http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/index.jsp
- IBM System Planning Tool website

  http://www.ibm.com/systems/support/tools/systemplanningtool/
- IBM Fix Central website

  http://www.ibm.com/support/fixcentral/
- Power Systems Capacity on Demand website

  http://www.ibm.com/systems/power/hardware/cod/
- Support for IBM Systems website

  http://www.ibm.com/support/entry/portal/Overview?brandind=Hardware~Systems~Power
- IBM Power Systems website

  http://www.ibm.com/systems/power/
- IBM Storage website

  http://www.ibm.com/systems/storage/
- IBM Systems Energy Estimator

  http://www-912.ibm.com/see/EnergyEstimator/
- Migration combinations of processor compatibility modes for active Partition Mobility

  http://publib.boulder.ibm.com/infocenter/powersys/v3r1m5/topic/p7hc3/iphc3pcmcombosact.htm

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

IBM Power System S822

Technical Overview and Introduction

Redpaper™

Outstanding performance based on POWER8 processor technology

2U scale-out rack-mount server

Improved RAS features

This IBM® Redpaper™ publication is a comprehensive guide covering the IBM Power System S822 (8284-22A) server that support IBM AIX®, and Linux operating systems. The objective of this paper is to introduce the major innovative Power S822 offerings and their relevant functions:

- The new IBM POWER8™ processor, available at frequencies of 3.42 GHz, and 3.89 GHz.
- Significantly strengthened cores and larger caches
- Two integrated memory controllers with improved latency and bandwidth
- Integrated I/O subsystem and hot-pluggable PCIe Gen3 I/O slots
- Improved reliability, serviceability, and availability functions
- IBM EnergyScale™ technology that provides features such as power trending, power-saving, capping of power, and thermal measurement.

This publication is for professionals who want to acquire a better understanding of IBM Power Systems™ products.

This paper expands the current set of IBM Power Systems documentation by providing a desktop reference that offers a detailed technical description of the Power S822 system.

This paper does not replace the latest marketing materials and configuration tools. It is intended as an additional source of information that, together with existing sources, can be used to enhance your knowledge of IBM server solutions.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

REDP-5102-00